ECAI 2023
K. Gal et al. (Eds.)

doi:10.3233/FAIA230608

Region-Specific Prototype Customization for Weakly Supervised Semantic Segmentation

Ruiguo Yu[a,b,c,d], **Yihang Zhao**[a,b,c], **Mei Yu**[a,b,c,d], **Jie Gao**[a,b,c], **Chenhan Wang**[e], **Ruixuan Zhang**[a,b,c] and **Xuewei Li**[a,b,c,d;*]

[a]College of Intelligence and Computing, Tianjin University, Tianjin, 300350, China.
[b]Tianjin Key Laboratory of Cognitive Computing and Application, Tianjin, 300350, China.
[c]Tianjin Key Laboratory of Advanced Networking, Tianjin, 300350, China.
[d]School of Future Technology, Tianjin University, Tianjin, 300350, China.
[e]OpenBayes (Tianjin) IT Co., Ltd., Tianjin, 300456, China.

Abstract. It is well known that weakly supervised semantic segmentation requires only image-level labels for training, which greatly reduces the annotation cost. In recent years, prototype-based approaches, which prove to substantially improve the segmentation performance, have been favored by a wide range of researchers. However, we are surprised to find that there are semantic gaps between different regions within the same object, hindering the optimization of prototypes, so the traditional prototypes can not adequately represent the entire object. Therefore, we propose region-specific prototypes to adaptively describe the regions themselves, which alleviate the effect of semantic gap by separately obtaining prototypes for different regions of an object. In addition, to obtain more representative region-specific prototypes, a plug-and-play Spatially Fused Attention Module is proposed for combining the spatial correlation and the scale correlation of hierarchical features. Extensive experiments are conducted on PASCAL VOC 2012 and MS COCO 2014, and the results show that our method achieves state-of-the-art performance using only image-level labels.

1 Introduction

The semantic segmentation task is one of the important tasks in the field of computer vision, which have been widely used in autonomous driving [10], remote sensing image interpretation [14] and medical imaging [28], etc. Thanks to the proposal of deep convolutional neural networks (CNNs), the semantic segmentation task in fully supervised domains has made leaps and bounds. However, one of the biggest problems in training models of fully supervised semantic segmentation task is that the pixel-level labels of the dataset depend on a large amount of human and time resources. To reduce the workload, weak labels are introduced for weakly supervised semantic segmentation (WSSS) tasks, such as bounding boxes [19], scribbles [21], points [4], and image-level labels [2], etc. Among the above labels, image-level labels have received a lot of favorable attention due to the fact that they consume less resources and are easier to acquire.

Most existing WSSS methods based on image-level labels are roughly divided into three steps: 1) firstly, the initial seed region is acquired through Class Activation Map (CAM) [36], 2) the initial seed region is then optimized by inter-pixel feature relationships, 3) finally, methods such as denseCRF [16] are used to refine the edges to obtain a refined pseudo mask. However, the initial seed region obtained through CAM has the problems of insufficient foreground activation and excessive background activation, which seriously affect the quality of object activation.

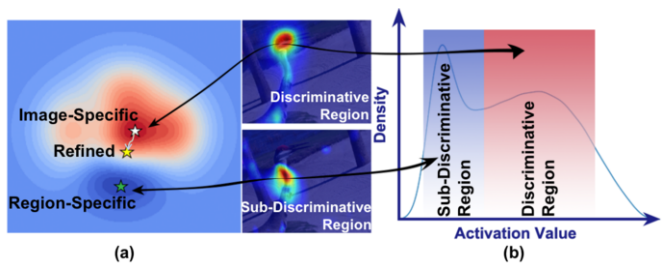

Figure 1. (a) Feature manifold of the discriminative region (red contour lines) and sub-discriminative region (blue contour lines). Our method adds an additional sub-discriminative region-specific prototype to the conventional image-specific prototype (i.e. discriminative region-specific prototype). (b) The distribution density of activation values within all object regions of PASCAL VOC 2012.

Most of existing studies focus on the ways to improve the quality of initial seed region acquired by CAM. Some studies try to obtain feature vectors that can represent the entire object region, and introduce a new metric space that provides additional supervised information based on CAM so that the quality of object activation can be improved. Therefore, prototypes [7, 23] are proposed, which are defined as representative embeddings of the classes. In the inference stage, the class to which each pixel belongs is determined by the similarity of that pixel feature to each class of prototypes. It follows that obtaining a more representative prototype is the key to improve the performance of prototype-based methods.

Current prototypes can be broadly classified into two types based on the acquisition method, one of which is generalized as class-specific prototype [23], where the models obtain one or several prototypes for each class of dataset during training. However, due to the diversity of images, using only a small number of prototypes is under-representative and does not guarantee the activation of all object regions of that class in the dataset. Therefore, efforts have been made to obtain the more representative prototypes to improve performance gains. Another prototype, the image-specific prototype in

* Corresponding Author. Email: lixuewei@tju.edu.cn

SIPE [7], comes into being, obtaining specific prototypes for each class of each image. Although the above prototype-based approaches can optimize the prototype representativeness and improve the quality of object activation to a certain extent, both class-specific prototype and image-specific prototype obtain representative prototypes through the features of the CAM of interest. We argue that there are two issues with the object features used for prototype calculation:

1. There are semantic gaps between different regions within the same object, according to Figure 1(a). As shown in Figure 1(b), we have statistically analyzed the distribution of activation values within all object regions of PASCAL VOC 2012, and the distribution shows two peaks, indicating that there are two kinds of feature distribution within the object regions (if the features are similar, the activation values should also be similar). Therefore, it can be assumed that the number of semantic gaps is 1. The refined prototype obtained by expanding the initial seed region could capture the features of sub-discriminative region. However, due to semantic gaps, the two region features are opposite to each other for the prototype's optimization direction, which does not guarantee that the prototype fully comprehends the semantic patterns of all regions within the object.

2. The current feature representation do not contain both spatial correlation and scale correlation at the same time, so object features are not global in nature. Where the lack of spatial correlation of features makes it difficult to capture long-distance feature associations within the image, and the lack of scale correlation of features makes it difficult to capture correlation between feature maps, leading to independent optimization of features in each layer. Although some existing methods [12, 13, 15, 20] have mentioned the above two correlations, the two correlations are still independent of each other.

To deal with the above issues, we propose that the obtained prototype can adequately represent the entire object as long as the prototype is unaffected by semantic gaps, and the features used for obtaining prototype can adequately represent pixel information. Therefore, in this paper, an effective model, called Region-Specific Prototype Customization (RPC), is proposed and experimentally validated on the Pascal VOC 2012 and MS COCO 2014 public dataset. The experiment results show that pseudo masks and segmentation results achieve state-of-the-art performance using only image-level labels. In summary, the main contributions of this paper are as follows:

- In order to alleviate the influence of semantic gaps between different regions and make the prototype fully comprehend the semantic patterns of all regions within the object, the Region-Specific Prototype is proposed to further adaptively describe the regions themselves, which obtains the most representative prototypes for discriminative and sub-discriminative regions of the object, respectively, on the basis of image-specific prototype.
- To further optimize the feature representation, the Spatially Fused Attention Module (SFA Module) is proposed. This module uses a multiplicative fusion strategy to simply and efficiently fuse the spatial correlation and scale correlation of hierarchical features for co-optimization, helping to obtain more representative region-specific prototypes. And it can be plug-and-play in other models.

2 Related Work

2.1 Visual Attention Module

Initially, channel-wise attention [15] was proposed to explicitly model the inter-dependencies between the channels of convolutional feature to improve the representation capability of the network. While in the field of semantic segmentation spatial correlation is significantly more important than channel correlation, so a series of spatial-wise attention was proposed [25, 29, 32, 35], which captures the long-distance correlation among feature maps by calculating the similarity between pixel features, driving the development of CNNs in computer vision. The recent trend is Transformer [3], which completely abandons network structures, such as RNNs and CNNs, and captures global correlation among sequences through self-attention, greatly improving network performance and having a profound impact in natural language processing and computer vision.

Compared to conventional spatial-wise attention, we additionally consider scale correlation based on spatial-wise attention to further model global correlation. Although the emerging self-attention in Transformer has the ability to capture long-distance correlation, it cannot take advantage of the priori knowledge of the image itself, such as scale, translation invariance, etc. This leads to the fact that the self-attention is only effective on the basis of large amount of data. In the case of small amount of data, the proposed CNN-based SFA module achieves better performance with fewer parameters.

2.2 Prototype-based Method

In earlier research, Cho et al. [8] use an approach similar to unsupervised learning in order to mine class-specific prototypes by clustering. Considering that the information interaction between images can be utilized, Liu et al. [23] propose RPNet to explore the diversity across image of the training set, using prototypes obtained from multiple images to reactivate unactivated regions. In order to make the class-specific prototypes adequately represent the feature representation of the corresponding class, memory bank-based method [24] and method based on contrast learning [9] are proposed, which significantly improve the performance. The prototypes obtained by the above methods represent the class centers of the entire dataset, resulting in unequal activation for different individual images. Therefore, image-specific prototype [7] is proposed to adaptively describe the image itself. However, since the region of the prototype is restricted by the local optimum problem of CAM, and there are semantic gaps between different regions within the object, the prototype obtained by the previous approaches cannot fully comprehends the semantic patterns of all regions within the object. In this circumstance, our region-specific prototype is born.

3 Preliminary

Class Activation Map. Most of WSSS methods are based on CAM to obtain the initial seed region. In particular, the input image is fed into a classification network, and the CAM $M = \{M_c\}_{c=1}^{C}$ over C foreground classes can be obtained as follows:

$$M_c = ReLU(\omega_c^{\tau} F), \forall c \in C \tag{1}$$

where ω_c is the classification weight of class c and F is the semantic feature from the last layer of the network.

Recently, there is an equivalent but simpler way to obtain the CAM, as shown in Equation 2, which replace the linear layer with the convolutional layer f. The features of the last layer F are fed directly into the convolutional layer and then $ReLU$ activation is performed to obtain the CAM. The final prediction scores for each class are obtained by global pooling CAM.

$$M = ReLU(f(\theta, F)) \tag{2}$$

Prototype. The obtained CAM is normalized and assigned as weights to the features at the corresponding pixel positions, and the

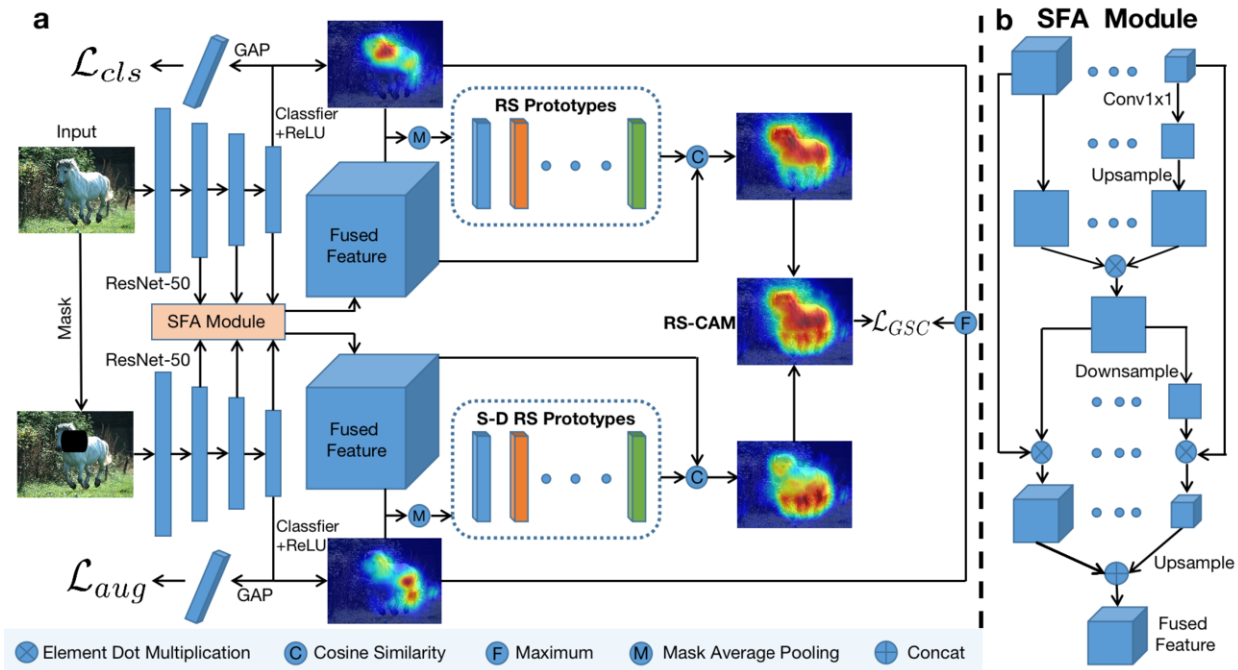

Figure 2. An overview of our method. There are two main components: (a) The architecture of RPC. (b) The schematic of SFA Module. In the figure, RS Prototypes denotes Region-Specific Prototypes of discriminative regions, S-D RS Prototypes denotes Region-Specific Prototypes of sub-discriminative regions, and RS-CAM denotes the final object activation generated by our method.

prototype is finally obtained by global weighted summation as follows:

$$P_c = \frac{\Sigma_{i\in\Omega} S_{i,c} V_i}{\Sigma_{i\in\Omega} S_{i,c}}, \forall c \in C \tag{3}$$

where C denotes the set of classes. P_c is the prototype representing class c. i denotes the location of each pixel in the pixel space Ω. $S_{i,c}$ denotes the activation score of class c for pixel at location i, and V_i denotes the feature vector at that location.

4 Method

4.1 Overall Framework

As mentioned in Section 1, we find that there are semantic gaps between different regions within the same object. Based on this finding, we propose an effective model to focus on the discriminative and sub-discriminative regions of the object, respectively, in order to describe the corresponding regions adaptively.

The overall framework of the proposed method can be found in Figure 2(a). It consists of three main components: a network focusing on discriminative regions, another network focusing on sub-discriminative regions, and an attention module (SFA Module). The backbones of the two networks can be any CNN classifiers, such as ResNet series, which are optimized with two classification loss functions, i.e. $\mathcal{L}_{cls}$ and $\mathcal{L}_{aug}$, so that they acquire the ability to activate the discriminative and sub-discriminative region within object. The two activation values are used as feature weights to obtain region-specific prototypes for the regions corresponding to the stages, respectively. The similarity between the two-stage pixel features and corresponding region-specific prototypes is calculated separately for pixel-level prediction. The final fusion of the predictions of the two stages is used as the final object activation RS-CAM. In addition, General-Specific Consistency Loss ($\mathcal{L}_{GSC}$) is utilized to minimize the gap between the RS-CAM and CAM. Therefore, the overall optimized loss function can be formulated as follows:

$$\mathcal{L} = \mathcal{L}_{cls} + \mathcal{L}_{aug} + \mathcal{L}_{GSC} \tag{4}$$

4.2 Region-Specific Prototype

Region-specific prototypes are proposed to adaptively represent the distribution of features of discriminative and sub-discriminative regions in the object respectively, alleviating the influence of semantic gaps and further activating complete and accurate object regions. To this end, an effective method (RPC) for region-specific prototype customization is designed, as shown in Figure 2(a). In the first stage, a region-specific prototype is obtained from discriminative region. In the second stage, another region-specific prototype is obtained from sub-discriminative region. Finally the two prototypes are used to co-predict the entire object. Specifically, a given image x is fed into backbone to obtain classification predictions and the CAM M. SFA Module is added to the backbone after making full use of the hierarchical features to obtain the fused feature F_s that contains the spatial attention of each layer. The input image x' of the second stage is obtained by setting a threshold value α for M to mask the discriminative region according to Equation 5:

$$x' = \mathbb{1}(max(M) < \alpha) * x \tag{5}$$

where $\mathbb{1}(\cdot)$ indicates that if the foreground activation score is less than α output 1, otherwise output 0. $max(\cdot)$ indicates taking the maximum value along the channel dimension. The threshold α is set to 0.7 in this paper. The mask is to erase the specific region of original image according to the high response region in CAM, and still retain the low-response region, instead of completely erasing the whole object (it cannot erase the whole object in fact). The processes performed on the masked image are the same as that at the first stage. Note that

stage 2 uses an additional backbone network that does not share parameters with stage 1. Then the second stage CAM M' is obtained by an additional classifier. The respective region-specific prototypes P_c, P'_c are customized using M, M' with the corresponding fused features F_s, F'_s as follows:

$$P_c = Avg\_pool(M_c * F_s), \forall c \in C \tag{6}$$

$$P'_c = Avg\_pool(M'_c * F'_s), \forall c \in C \tag{7}$$

where $Avg\_pool(\cdot)$ denotes global average pooling. Different from image-specific prototypes and class-specific prototypes, our method considers the semantic gaps between different regions within the same object and focus individually on discriminative and sub-discriminative regions. The independent optimization of region-specific prototypes acquired in two stages alleviates the influence of contradictory optimization directions caused by semantic gaps, ensuring the integrity of the corresponding region activations and simultaneously suppressing background activations.

The final object activation RS-CAM Q is calculated by integrating F_s, F'_s, P_c, P'_c as shown in Equation 8:

$$Q = max(Sim(F_s, P_c), Sim(F'_s, P'_c)) \tag{8}$$

where $Sim(\cdot)$ denotes the similarity calculation, and cosine similarity is used in this paper. Since the acquired region-specific prototypes fully understand the semantic patterns of the respective regions while the gap with the background features gets larger, the complete activation of the respective regions can be guaranteed and the background activation is suppressed to refine the object edges. Through the combination of the two predictions, the above advantages can be integrated to obtain a more completed and accurate object activation.

4.3 Spatially Fused Attention Module

The current feature representation do not contain both spatial correlation and scale correlation at the same time, resulting in features that are not global in nature. Given that 1) the collaborative use of hierarchical features can grasp the scale correlation and co-optimize the features, and 2) the semantic segmentation task is a pixel-level dense classification task, it is essential to grasp the spatial correlation. We propose a simple but effective SFA Module to simultaneously combine the spatial correlation and the scale correlation of hierarchical features, and to optimize feature representations, customizing more representative region-specific prototypes.

As shown in Figure 2(b), given the hierarchical features acquired by the backbone, the spatial attention maps of each layer are obtained in a simpler way to capture the spatial correlation. Specifically, as shown in Equation 9, the spatially fused attention map S is obtained by elemental dot multiplication of spatial attention of each layer according to the multiplicative fusion strategy. Next, it is assigned to the initial features, as shown in Equation 10, to obtain the optimized features, which are finally concatenated along the channel dimension as the final fused features F_s for subsequent calculation of region-specific prototypes.

$$S = \prod_k (ReLU(BN(f_k(\theta_k; F_k)))) \tag{9}$$

$$S'_k = S * F_k \tag{10}$$

where F_k denotes the k-th layer feature map, f_k denotes the convolution operation corresponding to the k-th layer. Note that SFA Module uses a multiplicative fusion strategy. Due to the spatial invariance of the convolution operation, the features at the same spatial location in each layer correspond to each other, so the attention of each layer can be fused by multiplying the corresponding elements. In contrast to the unfused and additive fusion strategies, the multiplicative fusion strategy can capture scale correlation and co-optimize the parameters of each layer [31]. In particular, according to the chain rules, the gradient of an individual layer is not independent, but interacts with the other layers. In this case, when one layer fails to capture a better representation, it can affect the optimization of the remaining layers, thus co-optimizing the parameters of each layer.

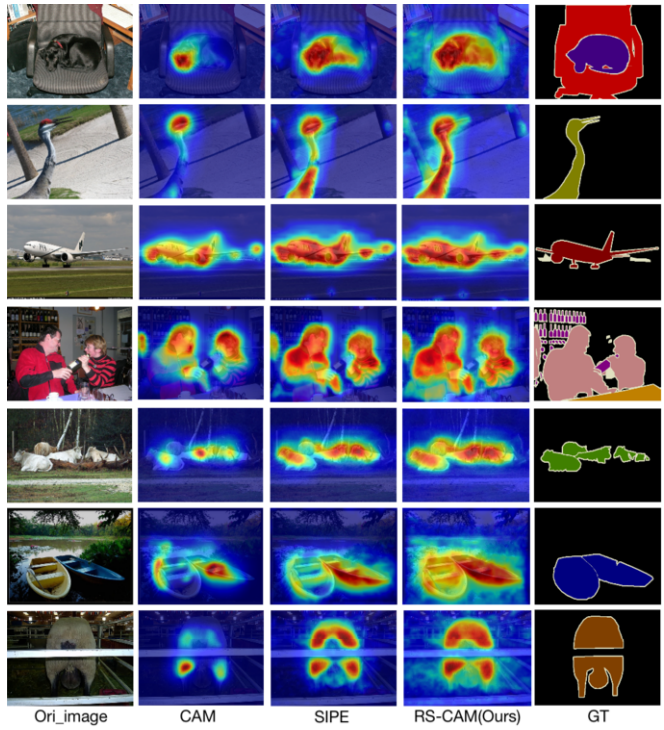

Figure 3. Comparison of object activations generated by various approaches on the PASCAL VOC 2012 public dataset.

Table 1. The mIoU (%) of pseudo masks generated by various approaches and that after refinement with denseCRF on the training set of the PASCAL VOC 2012.

Method	Pub.	Local.Maps	+denseCRF
SCE [5]	CVPR 20	50.9	55.3
SEAM [30]	CVPR 20	55.4	56.8
EDAM [33]	CVPR 21	52.8	58.2
AdvCAM [18]	CVPR 21	55.6	62.1
ECS [27]	ICCV 21	56.6	58.6
CSE [17]	ICCV 21	56.0	62.8
VWL-M [26]	IJCV 22	56.9	62.6
VWL-L [26]	IJCV 22	57.3	63.0
SIPE [7]	CVPR 22	58.6	64.7
RPC (Ours)		**60.5**	**70.5**

5 Experiments

In this section we introduce the details of our experiments. We first present the experimental settings and then compare our method with the state-of-the-art methods on the PASCAL VOC 2012 [11] and MS COCO 2014 [22]. Then series of ablation experiments are conducted

Table 2. The mIoU (%) of segmentation results on the Pascal VOC 2012 validation set and test set.

Method	Pub.	Backbone	Val	Test
SCE	CVPR 20	ResNet101	66.1	65.9
SEAM	CVPR 20	ResNet38	64.5	65.7
EDAM	CVPR 21	ResNet38	66.6	67.6
AdvCAM	CVPR 21	ResNet101	68.1	68.0
CSE	ICCV 21	ResNet38	68.4	68.2
RPNet	CVPR 22	ResNet101	68.0	68.2
SWP [34]	ICASSP 22	ResNet101	66.1	-
VWL-M	IJCV 22	ResNet101	68.7	69.2
VWL-L	IJCV 22	ResNet101	69.2	69.2
SIPE	CVPR 22	ResNet38	68.2	69.5
SIPE	CVPR 22	ResNet101	68.8	69.7
RPC (Ours)		**ResNet38**	**69.5**	**69.6**
RPC (Ours)		**ResNet101**	**70.7**	**71.2**

Table 3. The mIoU (%) of segmentation results on the MS COCO 2014 validation set.

Method	Pub.	Backbone	Val
SEAM	CVPR 20	ResNet38	31.9
CSE	ICCV 21	ResNet38	36.4
RPNet	CVPR 22	ResNet101	38.6
VWL-M	IJCV 22	ResNet101	36.1
VWL-L	IJCV 22	ResNet101	36.2
SIPE	CVPR 22	ResNet38	43.6
SIPE	CVPR 22	ResNet101	40.6
RPC (Ours)		**ResNet38**	**43.0**
RPC (Ours)		**ResNet101**	**44.5**

to demonstrate the effectiveness of the proposed method. And finally comparison experiments on the number of Region-Specific Prototypes, different combination of feature layers and the setting of mask threshold determine the optimal parameters for the model. Note that the mean intersection over union (mIoU) is used as a metric to evaluate segmentation results in the same way as other semantic segmentation methods and Local.Maps is the quality of the object localization map. The results for the PASCAL VOC 2012 test set are obtained from the official evaluation server. The experiments in sections 5.3 - 5.6 are conducted on the Pascal VOC 2012 dataset.

5.1 Experimental Settings

The proposed RPC model uses ResNet50 with ImageNet initialization as the backbone network, where the fully connected layers are replaced with convolutional layers. In the training stage, the model is trained with a batch size of 16, and for 5 epoches on RTX 3090, using SGD optimizer with a momentum of 0.9 and a weight decay of 1e-4. The initial learning rate is set to 0.1, and gradually decays with the use of the Poly strategy.

5.2 Comparison Experiment

5.2.1 Improvement of Pseudo Masks

A qualitative comparative analysis of the visualized object activations is performed as shown in Figure 3, showing that the object activations obtained by RPC are significantly more completed and accurate, with better activations in some sub-discriminative regions and finer boundaries. In addition, a quantitative comparison of model performance is performed using the mIoU metric. As shown in Table 1, RPC outperforms other approaches by improving the mIoU of the pseudo masks from 58.6% to 60.5%. The mIoU of the pseudo masks optimized with denseCRF is also improved from 64.7% to 70.5%. The above analyses demonstrate that the quality of the pseudo masks generated by RPC on the training set with denseCRF optimization reaches state-of-the-art performance and even outperforms the results of the remaining approaches using IRNet [1]. In contrast to other approaches, the performance improvement of RPC is greater when pseudo masks are optimized through denseCRF. Since it activates more object regions and the boundaries of the activation regions are closer to the ground truth, as shown in Figure 3, it is easier to achieve the best results with optimization. The above effect is attributed to the optimized feature representation by SFA Module, and to the alleviation of semantic gaps effects through region-specific prototype.

Table 4. Effect of the main contributions.

CAM	RS Prototype	SFA Module	Local.Maps (%)
✓			50.1
✓	✓		60.0
✓		✓	59.8
✓	✓	✓	60.5

Table 5. Comparison of different attention modules.

Method	Local.Maps (%)	Parameters
Spatial-wise attention	59.0	3.584K
Self-attention	58.2	6.878M
SFA Module (Ours)	**59.8**	**3.584K**

5.2.2 Improvement of Segmentation Results

To further validate the effectiveness of the model proposed in this paper, we train the Deeplabv2 [6] using the generated segmentation pseudo masks and compare the results with other state-of-the-art approaches on the validation and test set of Pascal VOC 2012, as well as the validation set of MS COCO 2014. The comparison of the segmentation results in Figure 4 shows that our method performs extremely well, close to the ground truth, benefiting from alleviating the effect of semantic gaps and fully exploring the entire region of the object. As shown in Table 2 and Table 3, using ResNet101 as the Deeplabv2 backbone, our method achieves 70.7% mIoU in the Pascal VOC 2012 validation set and 71.2% mIoU in the test set with only image-level labels, 1.5% higer than the result of SIPE. In addition, our method achieves 44.5% mIoU in the MS COCO 2014 validation set, 0.9% higer than the result of SIPE. The above results verify the effectiveness of our RPC model.

5.3 Ablation Experiments

5.3.1 Region-Specific Prototype

The ablation experiment is conducted on the region-specific prototype, as shown in Table 4, showing that the mIoU of object activation can be improved to 60.0%. The activations of two stages are visualized and compared, as shown in Figure 5. Two benefits of region-specific prototypes can be identified: 1) At each stage, the respective regions are fully explored due to mitigating the influence of semantic gaps. As in the case of the horse in the figure, the head and body

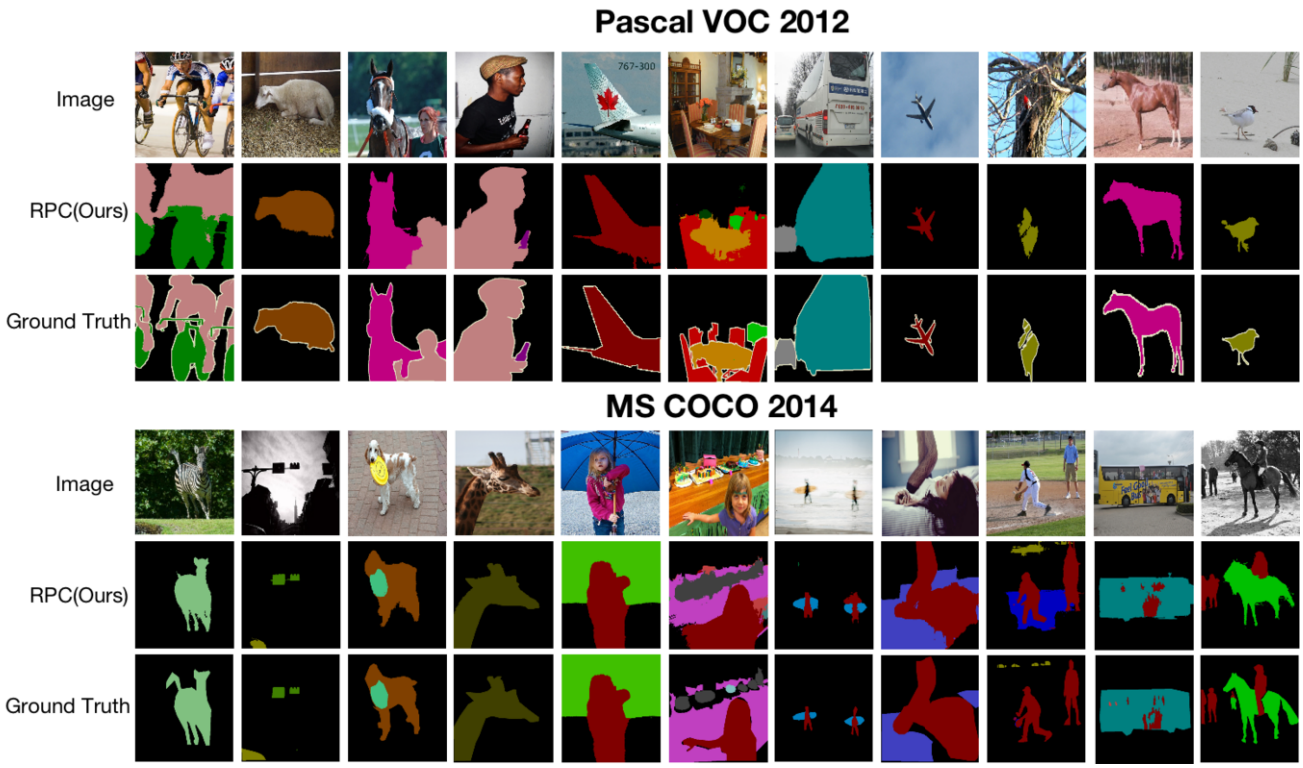

Figure 4. Segmentation results on the PASCAL VOC 2012 and MS COCO 2014 validation sets.

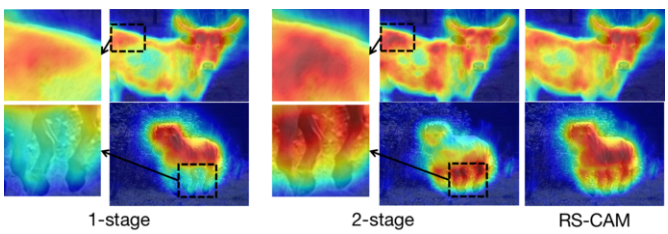

Figure 5. Comparison of object activations in each stage of RPC.

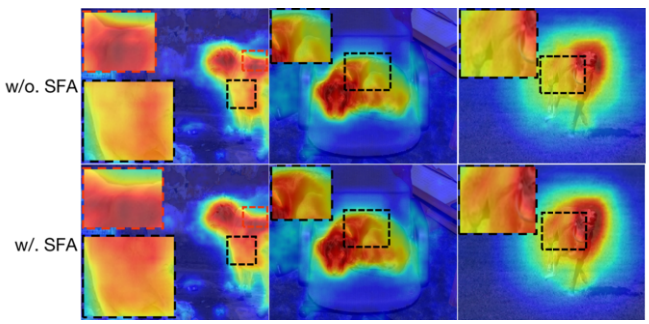

Figure 6. Comparison of object activations for removing and adding SFA Module.

parts are activated in the first stage, the limbs are activated in the second stage, and the final object activation obtained by fusion covers the entire object region. 2) Since the prototype adaptively describe the region itself, it is more representative, thus suppressing the background activation and making the edge activation more refined. In summary, the proposed region-specific prototype alleviates the influence of semantic gaps and improves the quality of object activation.

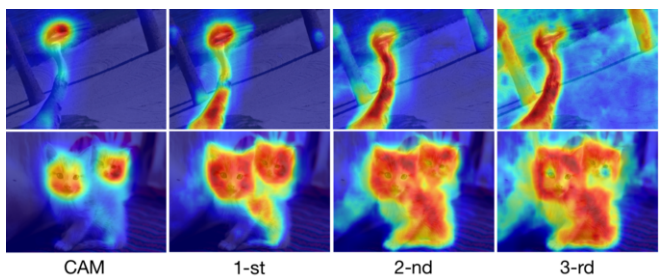

Figure 7. Object activation of the k-th prototype (for comparison, we have added the 3-rd region-specific prototype).

5.3.2 SFA Module

In order to optimize the feature representation, the SFA module is introduced to co-optimize the features of each layer. As shown in Table 4, our SFA module can improve the mIoU score up to 60.5% on the basis of region-specific prototype. In addition, a visualization comparison analysis of SFA Module in Figure 6 shows that SFA Module combines the spatial correlation and the scale correlation of hierarchical features and customize more representative region-specific prototypes, allowing full exploration of structurally similar regions (e.g., body parts of sheep). It also suppresses false activation of the background, enabling finer prediction of edges (e.g., back edges of sheep). To further demonstrate the advancement of SFA Module over existing attention mechanisms, such as spatial-wise attention and self-attention in Transformer, we conducted the comparison experiments in Table 5. Compared to spatial-wise attention, SFA Module improves the quality of object activation by 0.8% without increasing the number of parameters. Although the number of parameters in SFA Module is only 0.05% of that in self-attention, SFA Mod-

Table 6. Comparison experiment of adding SFA Module or region-specific prototype to different baselines.

Baseline	Original (%)	+SFA Module (%)	+RS Prototype (%)
RPNet	50.8	53.2 (2.4 ↑)	53.7 (2.9↑)
SIPE	58.6	59.8 (1.2 ↑)	60.0 (1.4↑)

Table 7. Comparison experiment of the number of Region-Specific Prototypes.

The number of RS Prototypes	Local.Maps (%)
1	58.6
2	**60.0**
3	58.0

ule improves the quality of object activation by 1.6%. The above experiment results show that our SFA Module could achieve better performance with fewer parameters.

5.3.3 *Plug-and-play*

To further validate the effectiveness of the region-specific prototype and the plug-and-play performance of SFA Module, we conduct comparison experiments on two baselines, such as RPNet [23] and SIPE [7], as shown in Table 6. The quality of object activation is improved after inserting SFA Module or Region-Specific Prototype, and the mIoU is improved by 2.4%, 2.9% on RPNet and 1.2%, 1.4% on SIPE, respectively. The above experiment results show that not only the SFA module can be introduced to any network, but also region-specific prototype can optimize other prototype-based model.

5.4 *The Optimal Number of Region-Specific Prototypes*

The optimal number of region-specific prototypes is determined by conducting comparative experiments for region-specific prototype, as shown in Table 7. The best performance of 60.0% is achieved when the number of region-specific prototypes is 2. To facilitate understanding, the object activation for the k-th ($k = 1, 2, 3$) prototypes is visualized in Figure 7. It is obvious that the 2-nd region-specific prototypes can further activate sub-discriminative regions compared to the first, while the 3-rd region-specific prototypes over-activate the background without significantly improving the activation of the second. It is can be explained as follows: when the number is 1, a single region-specific prototype cannot fully comprehend the semantic patterns of all regions within the object due to the semantic gap, resulting in insufficient activation. While when the number is 3, the 3-rd prototype will contain background noise owing to the uncertainty of the erasure method, which will over-activates the background, leading to the decrease in performance.

Therefore, combining the above analysis with Figure 1, we can determine the number of semantic gaps within the object as 1 and set the number of region-specific prototypes to 2, thus conforming to the cognitive model of regional features and alleviating the influence of semantic gaps to achieve the best performance.

Table 8. The comparison experiments of the feature layers selected by SFA module.

layer1	layer2	layer3	layer4	Local.Maps (%)
			✓	58.7
		✓	✓	58.8
	✓	✓	✓	**59.8**
✓	✓	✓	✓	59.0

Table 9. Comparison experiments of mask threshold setting.

α	0.3	0.4	0.5	0.6	**0.7**	0.8	0.9
Local.Maps (%)	57.0	58.0	58.8	59.8	**60.0**	59.9	59.7

5.5 *Effective Combination of Feature Layers in SFA Module*

A comparative experiment is designed for SFA Module to verify which combination of features is the most suitable. As shown in Table 8, the best performance is achieved by combining layer 2-4, with an mIoU metric of 59.8%, outperforming other feature combinations such as layer 4, layer 3-4, and layer 1-4. We explain this by the fact that the features of layer 1 contain more background noise than the features of the other layers, so that the spatial attention obtained from them is less correct, which will introduce errors in the feature fusion process and lead to hindering the optimization of the features of the remaining layers. When using features combined from layer 2-4, the semantic information contained in the deep features and the structural information contained in the shallow features can be optimally fused so that the feature representation can be optimized to the maximum extent.

5.6 *Mask Threshold*

In order to explore the effect of mask threshold α as a hyperparameter, we perform comparison experiments on various threshold settings for region-specific prototype, as shown in Table 9. When the mask threshold is set to 0.7, its object activation has the highest quality of 60.0%. Furthermore, it is evident that the proposed region-specific prototype is not sensitive to the threshold value, that is, when α varies in the range of 0.6 to 0.9, the model performance only changes slightly. The above results show that the effectiveness of region-specific prototype is not limited by the setting of α, which is a significant reason why the region-specific prototype can be directly used to optimize other prototype-based methods.

6 Conclusion

In this paper, the RPC model is proposed in the field of WSSS. The region-specific prototypes are proposed to adaptively describe discriminative and sub-discriminative regions themselves respectively, which alleviates the effect of semantic gaps and motivates the model to fully explore the corresponding regions so as to explore integral object. To obtain more representative region-specific prototypes, the SFA Module is additionally introduced to optimize the feature representation. Extensive experiments have been conducted for validation, and the results show that RPC achieves state-of-the-art performance with only image-level labels.

References

[1] Jiwoon Ahn, Sunghyun Cho, and Suha Kwak, 'Weakly supervised learning of instance segmentation with inter-pixel relations', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2209–2218, (2019).

[2] Jiwoon Ahn and Suha Kwak, 'Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4981–4990, (2018).

[3] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei, 'Beit: Bert pre-training of image transformers', *arXiv preprint arXiv:2106.08254*, (2021).

[4] Amy Bearman, Olga Russakovsky, Vittorio Ferrari, and Li Fei-Fei, 'What's the point: Semantic segmentation with point supervision', in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14*, pp. 549–565. Springer, (2016).

[5] Yu-Ting Chang, Qiaosong Wang, Wei-Chih Hung, Robinson Piramuthu, Yi-Hsuan Tsai, and Ming-Hsuan Yang, 'Weakly-supervised semantic segmentation via sub-category exploration', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8991–9000, (2020).

[6] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille, 'Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **40**(4), 834–848, (2017).

[7] Qi Chen, Lingxiao Yang, Jian-Huang Lai, and Xiaohua Xie, 'Self-supervised image-specific prototype exploration for weakly supervised semantic segmentation', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4288–4298, (2022).

[8] Jang Hyun Cho, Utkarsh Mall, Kavita Bala, and Bharath Hariharan, 'Picie: Unsupervised semantic segmentation using invariance and equivariance in clustering', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16794–16804, (2021).

[9] Ye Du, Zehua Fu, Qingjie Liu, and Yunhong Wang, 'Weakly supervised semantic segmentation by pixel-to-prototype contrast', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4320–4329, (2022).

[10] Andreas Ess, Tobias Müller, Helmut Grabner, and Luc Van Gool, 'Segmentation-based urban traffic scene understanding.', in *BMVC*, volume 1, p. 2. Citeseer, (2009).

[11] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman, 'The pascal visual object classes challenge: A retrospective', *International Journal of Computer Vision*, **111**, 98–136, (2015).

[12] Chengling Gao, Hailiang Ye, Feilong Cao, Chenglin Wen, and Feng Zhang, 'Multiscale fused network with additive channel-spatial attention for image segmentation', *Knowledge-Based Systems*, **214**(8), 106754, (2021).

[13] Pengcheng Guo, Xiangdong Su, Haoran Zhang, and Feilong Bao, 'Mcdalnet: Multi-scale contextual dual attention learning network for medical image segmentation', in *2021 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, (2021).

[14] Mohammad D Hossain and Dongmei Chen, 'Segmentation for object-based image analysis (obia): A review of algorithms and challenges from remote sensing perspective', *ISPRS Journal of Photogrammetry and Remote Sensing*, **150**, 115–134, (2019).

[15] Jie Hu, Li Shen, and Gang Sun, 'Squeeze-and-excitation networks', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7132–7141, (2018).

[16] Philipp Krähenbühl and Vladlen Koltun, 'Efficient inference in fully connected crfs with gaussian edge potentials', *Advances in Neural Information Processing Systems*, **24**, (2011).

[17] Hyeokjun Kweon, Sung-Hoon Yoon, Hyeonseong Kim, Daehee Park, and Kuk-Jin Yoon, 'Unlocking the potential of ordinary classifier: Class-specific adversarial erasing framework for weakly supervised semantic segmentation', in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6994–7003, (2021).

[18] Jungbeom Lee, Eunji Kim, and Sungroh Yoon, 'Anti-adversarially manipulated attributions for weakly and semi-supervised semantic segmentation', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4071–4080, (2021).

[19] Jungbeom Lee, Jihun Yi, Chaehun Shin, and Sungroh Yoon, 'Bbam: Bounding box attribution map for weakly supervised semantic and instance segmentation', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2643–2652, (2021).

[20] Haixing Li, Haibo Luo, Wang Huan, Zelin Shi, Chongnan Yan, Lanbo Wang, Yueming Mu, and Yunpeng Liu, 'Automatic lumbar spinal mri image segmentation with a multi-scale attention network', *Neural Computing and Applications*, **33**, 11589–11602, (2021).

[21] Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun, 'Scribblesup: Scribble-supervised convolutional networks for semantic segmentation', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3159–3167, (2016).

[22] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick, 'Microsoft coco: Common objects in context', in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, (2014).

[23] Weide Liu, Xiangfei Kong, Tzu-Yi Hung, and Guosheng Lin, 'Cross-image region mining with region prototypical network for weakly supervised segmentation', *IEEE Transactions on Multimedia*, (2021).

[24] Weide Liu, Xiangfei Kong, Tzu-Yi Hung, and Guosheng Lin, 'Cross-image region mining with region prototypical network for weakly supervised segmentation', *IEEE Transactions on Multimedia*, (2021).

[25] Jongchan Park, Sanghyun Woo, Joon-Young Lee, and In So Kweon, 'Bam: Bottleneck attention module', *arXiv preprint arXiv:1807.06514*, (2018).

[26] Lixiang Ru, Bo Du, Yibing Zhan, and Chen Wu, 'Weakly-supervised semantic segmentation with visual words learning and hybrid pooling', *International Journal of Computer Vision*, **130**(4), 1127–1144, (2022).

[27] Kunyang Sun, Haoqing Shi, Zhengming Zhang, and Yongming Huang, 'Ecs-net: Improving weakly supervised semantic segmentation by using connections between class activation maps', in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7283–7292, (2021).

[28] Nima Tajbakhsh, Laura Jeyaseelan, Qian Li, Jeffrey N Chiang, Zhihao Wu, and Xiaowei Ding, 'Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation', *Medical Image Analysis*, **63**, 101693, (2020).

[29] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He, 'Non-local neural networks', in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7794–7803, (2018).

[30] Yude Wang, Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen, 'Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12275–12284, (2020).

[31] Jun Wei, Qin Wang, Zhen Li, Sheng Wang, S Kevin Zhou, and Shuguang Cui, 'Shallow feature matters for weakly supervised object localization', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5993–6001, (2021).

[32] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon, 'Cbam: Convolutional block attention module', in *Proceedings of the European conference on computer vision (ECCV)*, pp. 3–19, (2018).

[33] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16765–16774, (2021).

[34] Zhaozhi Xie and Hongtao Lu, 'Exploring category consistency for weakly supervised semantic segmentation', in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2609–2613. IEEE, (2022).

[35] Hengshuang Zhao, Yi Zhang, Shu Liu, Jianping Shi, Chen Change Loy, Dahua Lin, and Jiaya Jia, 'Psanet: Point-wise spatial attention network for scene parsing', in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 267–283, (2018).

[36] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba, 'Learning deep features for discriminative localization', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2921–2929, (2016).