Cross-Image Region Mining with Region Prototypical Network for Weakly Supervised Segmentation

Weide Liu, Xiangfei Kong, Tzu-Yi Hung, Guosheng Lin,

***Abstract*—Weakly supervised image segmentation trained with image-level labels usually suffers from inaccurate coverage of object areas during the generation of the pseudo groundtruth. This is because the object activation maps are trained with the classification objective and lack the ability to generalize. To improve the generality of the objective activation maps, we propose a region prototypical network (RPNet) to explore the cross-image object diversity of the training set. Similar object parts across images are identified via region feature comparison. Object confidence is propagated between regions to discover new object areas while background regions are suppressed. Experiments show that the proposed method generates more complete and accurate pseudo object masks, while achieving state-of-the-art performance on PASCAL VOC 2012 and MS COCO. In addition, we investigate the robustness of the proposed method on reduced training sets.**

***Index Terms*—weakly-supervised, segmentation**

I. Introduction

SEMANTIC segmentation is a task to assign each pixel in a scene image with a semantic category. One of the biggest challenges of this task is the numerous, time-prohibitive efforts entailed in the manual labeling of a large set of training images. To alleviate, or even free researchers from high costs of laborious annotations, more attention has been paid to weakly supervised semantic segmentation, in which annotations can be performed in a much-eased manner: rather than associating all pixels from an image with a label, weaker supervisions on the training images such as bounding boxes [1], [2], scribbles [3], image labels [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], and points [18] can be used as substitute annotations to train.

Compared with other annotations, the image-level label is not only cost-friendly but also readily accessible to the public. There is a broad choice of large image corpora with image-level labels such as ImageNet [19], PASCAL VOC [20], and images retrieved by search engines with their keywords as labels [21]. However, image-level labels merely provide categorical cues without any shape/texture information of any objects. It is crucial to utilize such cues to automatically produce coherent regions of objects with their pixel-level categorical information. To this end, Class Activation Maps (CAM) [22] has been wildly used to explore the localization and shape cues from image-level labels [4], [5], [6], [7], [8], [9], [10], [11], [12]. However, there is no guarantee of the completeness of the object activation regions of CAM, since they are trained only with the classification objective. Incomplete or wrong activation maps obtained by CAM may lead to sub-optimal performance.

To improve the quality of the object activation maps, many approaches are proposed to use the results from CAM as the seed activation regions and gradually refine their quality. Ahn *et al.* [4] use low-level image features to enhance the coherence of the activation maps. Wang *et al.* [12] expand the seed object regions by fusing the object activation maps computed on different spatial scales of the input image. Object regions that are not activated by CAM at one scale could be re-activated at a coarser level. However, this method does not explore the rich cross-image context within the same class in a training set. CIAN [23] addresses the cross-image context by computing an affinity map for each pixel of an image to another. However, the pixel of the query may not be the foreground and could end up retrieving the wrongly activated backgrounds. Hu *et al.* [24] find the cross-image context leveraging two self-erasing networks with hiding and erasing strategy to explore more object regions. New regions will be activated by the classification objective if the old ones are masked. However, this method cannot automatically stop the erasing even after the entire foreground is masked, resulting in the undesirable background being activated.

Inspired by the previous works, we propose that the key to improve the quality of the object activation maps is 1) a method to locate more complete object regions and 2) a mechanism to suppress wrongly activated background regions. We hold an assumption that the object regions possess significantly less diversity compared with the background. We use it to address the two key points: To point 1), we find that multiple training samples of the same class provide a diversity of object regions, see Figure 1. One activated object region in an image is helpful to explore more similar regions in another image of the same class. In particular, we generate the region prototypes to capture the diversity by collecting regions with similar features across the same object category images; then, we locate the inactivated object regions in these images by comparing their feature maps to the region prototypes. Similar

W. Liu is with School of Computer Science and Engineering, Nanyang Technological University (NTU), Singapore 639798 (e-mail: weide001@e.ntu.edu.sg).

X. Kong is with Ant Group, Singapore (e-mail: xiangfei.kong@antgroup.com).

T. Hung is with Delta Research Center, Singapore (e-mail: tzuyi.hung@deltaww.com).

G. Lin is with School of Computer Science and Engineering, Nanyang Technological University (NTU), Singapore 639798 (e-mail: gslin@ntu.edu.sg).

Corresponding author: Guosheng Lin.

regions are highlighted and assigned with large confidence values as new object regions.

To point 2), if an activated region belongs to the background, it is difficult to propagate its information among others. Since there is more diversity among backgrounds, fewer compared regions during prototype voting resemble the activated background region, and its information propagation is minimized. Our region comparison serves as a voting mechanism to assign them with lower confidence scores.

With the region prototype mechanism on cross-images, our proposed network generates better object activation maps to train a segmentation model. We conduct extensive experiments on PASCAL VOC 2012 [20] and MS COCO [25] to validate the effectiveness of our network. However, these datasets both have abundant training samples and do not cover a practical scenario in which fewer training samples are available. A good method should produce a high-quality pseudo groundtruth even when the training samples are scarce. Thus, we investigate the robustness of the proposed method on datasets of reduced training samples (*e.g.* up to $1/16$ of the number of training samples in PASCAL). Our main contributions are summarized as follows:

- The proposed extraction of region prototypes and their equal treatments ensure the reactivation of rare object regions that are not activated by CAM while preventing them from being dominated by prevalent features.
- The proposed voting mechanism on prototypes propagates important object information while suppressing backgrounds simultaneously, bringing in more robustness to either abundant or limited training images.
- We are the first to evaluate weakly supervised segmentation methods under limited training samples and in different backbones. The proposed method exhibits promising results under both conditions.
- We improve the performances on PASCAL VOC 2012 and MS COCO dataset and achieve new state-of-the-art.

II. Related Work

A. Semantic segmentation

Semantic segmentation is a fundamental computer vision task that assigns each pixel in the image with a category. Currently, state-of-the-art methods handle image semantic segmentation as a dense prediction task and adopt fully convolutional networks to make predictions [26], [27]. To make pixel-level dense predictions, encoder-decoder structures [28], [29], [30], [31], [17], [32], [33], [34], [35], [36], [37], [38], [39] are widely used to reconstruct high-resolution prediction maps. Typically an encoder gradually downsamples the feature maps, aiming to acquire large field-of-view and capture the semantic object information. Then, the decoder gradually recovers the fine-grained information. The field-of-view information is important for semantic segmentation tasks. Dilation connections [40] are often used to increase the field-of-view and then fuse high-level and low-level features for better predictions. We also follow the encoder-decoder design in our network and opt to transfer the guidance information in the low-resolution maps and use decoders to recover details.

B. Weakly supervised semantic segmentation.

To alleviate the data deficiency problem in image segmentation task, the weak supervision have been explored, such as supervision with bounding boxes [1], [2], scribbles [3], and image labels [4], [5], [6], [7], [8], [9], [10], [11], [13], [14], [15], [12], [41]. The weak supervision methods have the nature of less expensive to annotate and easily acquired. Previous state-of-the-art methods [4], [5], [42] mostly adopt CAMs [22] to localize the objects. However, the CAMs can only locate the most discriminative object regions, which is insufficient to train a segmentation model. Researchers have designed various ways to propagate the seed regions to the entire object to explore more object-relative regions. Hou *et al.* [24] expand the seed regions by erasing the currently discovered regions with prohibit attention and re-train the erased image. SPM [43] expands the seed regions with a discrepancy and intersection loss. Ahn *et al.* [4] propagate the seed regions to reach the boundary and utilize the inter-pixel relationship to refine the final pseudo ground mask. CIAN [23] retrieves the cross-image relation by calculating the dot product between every pixel from different images, which do not distinguish the foreground and background. In contrast to CIAN, we generate the region prototypes only from the confident object regions. CONTA [44] proposes a structural causal model to remove the confounding bias in image-level classification. Our PRNet also can suppress the wrong foreground by a voting mechanism with multiple prototypes. Compared with previous methods, our method expands the seed region by leveraging the cross-image relations from a novel region prototype perspective to mine the inactive object regions and suppress the dislocated foreground. [12] learns a network to enforce the classifier to recognize the common semantics from co-attentive objects. The cues of the foreground sometimes have to be refined by extra saliency information. In contrast, the proposed RPNet selects the confident object regions from different images to retrieve similar object parts while the background features can be suppressed simultaneously. [45], and the proposed method both seek to reactivate object regions of the same object category in different images. However, this work maintains a memory bank, and the centers of prototypes are used, which could cause the prototypes to be dominated by some most popular object regions, e.g., memory centers could be severely biased to dog heads, rather than dogs' whole bodies. This is perhaps less serious for detection tasks (which is the focus of [45]) but is very harmful to the image segmentation.

III. Proposed Method

A. Motivation

The object activation maps of CAM are sub-optimal for fully supervised segmentation training because they are trained via a classification objective. We argue that such maps suffer from two types of flaws: 1)incomplete object regions (foreground) and 2) falsely activated cluttered regions (background). Generally, when the training set has abundant samples in each category, the trained model benefits from it and tends to be robust to backgrounds. However, at the same time, the activated object region tends to be more incomplete since some

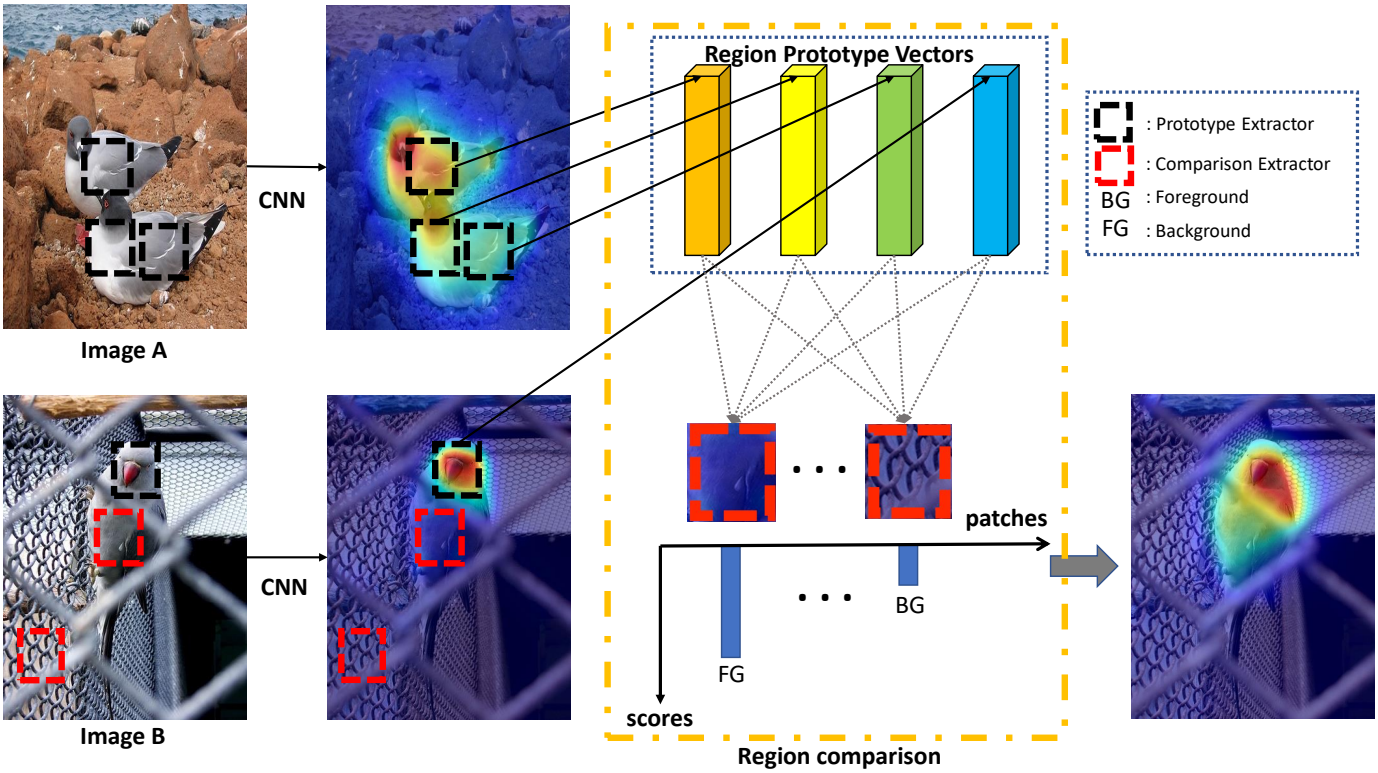

Fig. 1: Motivation and Region comparison. Given a pair of sample images that contain common classes as the network input, we first encoder the images into object activation maps, and generate the region prototype vectors by collecting the confident object regions. By comparing the similarity to the region prototype vectors, we re-weight the inactive object regions.

less distinctive parts tend to be suppressed by the classification objective, *e.g.* the torso part of a mammal which is less distinctive compared with its head. This means flaw 1) is more likely to happen. When the training samples become scarce, flaw 2) happens more as some backgrounds could be discriminative with fewer data available to counter possible overfitting. See Figure 4 and our observation in the experiment part in Sec. IV-D on truncated versions of *PASCAL VOC 2012*. It is crucial for a method to be robust to both abundant or scarce training samples and address the two mentioned flaws. To our knowledge, we are the first to investigate the number of training samples and handle both flaws in an elegant manner to produce high-quality object activation maps. [12],[23] and [24] consider the flaw 1) only, while [44] considers flaw 2) only.

B. Method Pipeline

As shown in Figure 2, given a training set of images with image-level labels, we train a model that automatically produces initial object regions. The model extracts object features similar to CAM [22] but is from multiple spatial resolutions. Based on these features, we explore the diversity of the training images in each object class: at each iteration, we compute a group of cross-image region prototype vectors on a pair of sample images that contain a common object class. These prototypes are used to compare with the feature maps of the inputs and find out more inactive foreground regions. Two loss functions, namely classification loss, and self-supervised loss, are used to ensure the precision of the foreground activation and suppress the background regions during training. The trained model is used to predict maps of initial object regions on the input images with their image-level labels.

The proposed method can be integrated into the popular pipeline for weakly supervised segmentation training [4], [5], [6], [44], [12]: After the initial object regions are computed, boundary refinement [4] is applied on them to generate the pseudo groundtruth, on which the fully-supervised segmentation model can be trained.

C. Revisiting Class Activation Maps

The proposed method starts with the results of CAM [22] and explores useful cross-image context based on them. CAM computes the object activation maps via a classification network backbone connected with a global average pooling (**GAP**) layer as follows:

$$O(x,y) = g(\frac{\Theta^T f_n(x,y)}{max \Theta^T f_n(x,y)}), \tag{1}$$

$$CAM(x,y) = O(x,y) \cdot c, \tag{2}$$

where f_n denotes the feature maps extracted from the last convolution block. (x,y) denote 2D coordinates, Θ is the classification weights and g is the is the activation function, which is $ReLU$ in this paper. The spatially normalized object activation map is denoted as $O(x,y)$ and the per-class object

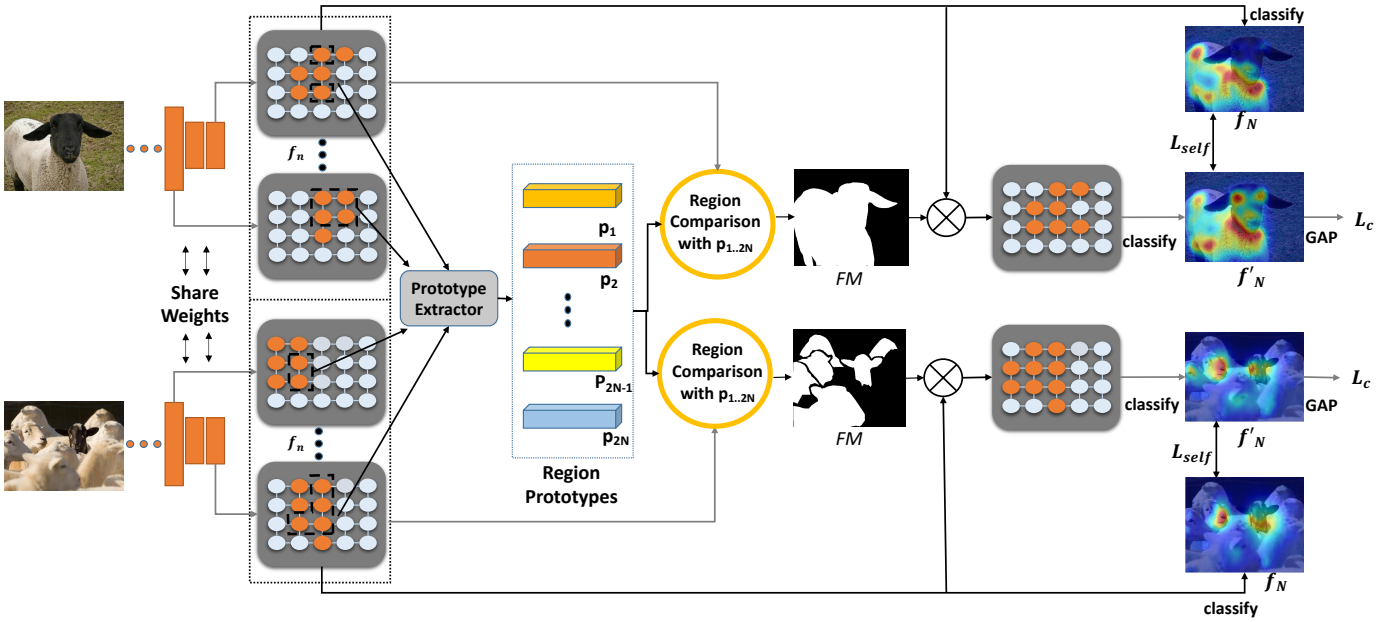

Fig. 2: The architecture of our methods. Given a pair of sample images that contain common classes as the network input, we first generate object activation maps with a parameter-shared encoder, then we reinforce the representations on the object-related regions, which are explored by region prototype vectors. We select different object-related regions of multiple-level features to generate the region prototype vectors and use them to generate the foreground probability maps. Besides the standard multi-label classification loss $\mathcal{L}_c$, we also add an auxiliary loss $\mathcal{L}_{self}$ that uses the refined prediction to supervise the training of the original prediction.

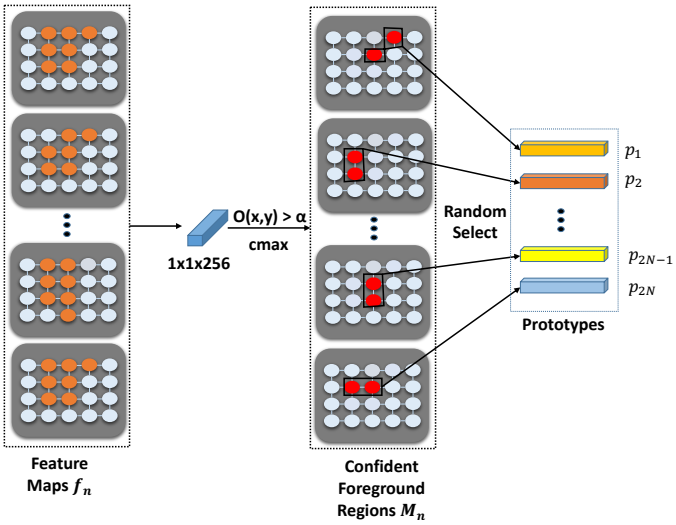

Fig. 3: The architecture of our prototype extractor

activation map $CAM(x, y)$ is computed by masking O with the image-level label vector c.

We follow the idea in [12] to fully explore the object information and compute O on multiple resolution feature maps, *e.g.* $\{f_n|n = 1...N\}$, where N is the number of blocks in a classification network backbone.

D. Region Prototypical Network

We propose the region prototype that captures distinctive features among sample images of the same object category: prototypes extracted from different samples are compared with their feature maps. We define this step as **region comparison** as shown in Figure 1. We hold a simple assumption: the diversity of the object regions is **far less** than the diversity of background. Thus, if an extracted prototype is truly an object region, with the help of the diversity of the training images, this prototype's information will propagate well during training and re-activate similar regions from other samples. On the contrary, if the prototype is background wrongly activated by the classification objective, the information propagation is much slower than object regions: the diversity of background is significantly more, and these prototypes can hardly find their similar neighbors. As the example shown in Figure 1, only the bird's head is activated in Image B, while the head and body regions are activated in Image A. The body features in image A can be used to guide image B to find out its inactivated body regions and highlight them. Based on this finding, we build a region prototype network to extract cross-image object features as prototypes, and takes pairs of images within the same object category as inputs. This aims to have full utilization of the object diversity among the training set. We are not the first to use image pairs as inputs: [23] explores cross-image context by computing an affinity map for each pixel for image pairs.

Generate regions prototype vectors. We follow the idea in [12] to fully explore the object information and compute O on multiple resolution feature maps of the image pair, *e.g.* $\{f_n(x, y)|n = 1...N\}$, where N is the index of resnet blocks in a classification network backbone. We unify of the number of channels by concatenation of each f_n with a 256 channel

1×1 convolution layer:

$$O_n(x,y) = conv * g(\frac{f_n(x,y)}{max f_n(x,y)}). \tag{3}$$

Unlike the prototypical network [46] which spatially averages the entire feature maps as the prototypes and introduces noise, our RPNet carefully selects activation regions as the region prototype vectors with high confidence only. Specifically, given a set of object activation maps $\{O_n(x,y)\}$, we compute their confident object activation maps by performing a maximum operation over the spatial dimension ($cmax$):

$$O_n^{'}(x,y) = cmax(O_n(x,y)). \tag{4}$$

In equation 3, We do not use the image-level label c as CAM does in equation 1 to locate object regions, due to the inherent flaw of the classification objective previously discussed. Instead, all channels are given equal possibilities to be highlighted as the foreground to produce more complete object regions. We generate the region prototype vectors $\{p_n\}$ as

$$p_n = \frac{\sum_{x,y} f_n(x,y) M_n(x,y)}{\sum_{x,y} M_n(x,y)}. \tag{5}$$

where

$$M_n(x,y) = \begin{cases} \mathrm{X}, & \text{if } O_n^{'}(x,y) > \alpha, \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

and X follows Bernoulli distribution $X \sim B(x,\alpha)$ and is equal to 1 with a probability β. For each f_n, we produce a binary confidence mask $M_n(x,y)$ by thresholding its corresponding confident object activation map $O^{'}$ with α, as in equation 6. The feature f_n is then masked by M_n to produce a prototype vector p_n by average pooling at the spatial dimension with normalization, see equation 5. The parameter X helps randomly discard some of the activated regions from M_n during training, and the location of the corresponding f_n will be masked. This is inspired by [24] to introduce robustness and help find more object regions. The effectiveness of discarding the activated regions is validated with experiments in Sec. IV-C. The complete flow of prototype generation is illustrated in Figure 3.

Region comparison. With the region prototype vectors, we try to filter the unconfident object-related regions in a non-parametric metric learning manner: at first, we calculate the similarity s_n between the feature maps of the last CNN block f_N and the region prototype vectors p_n, where $n = 1...2N$ at each spatial position; then, we find the maximum similarity and assign every spatial position with a probability to object foreground depending on it. Collecting all the max similarity values as a set, we form the foreground probability maps FM. We follow Oreshkin *et al.* [46] to use cosine as our similarity function, but we do not utilize the scale factor. The foreground probability maps FM are generated with the following steps:

$$s_n(x,y) = cos(p_n, f_N(x,y)), n = 1...2N, \tag{7}$$

$$FM(x,y) = max(s_n(x,y)), \tag{8}$$

where N is the number of feature maps extracted from different blocks of the CNN network.

We discuss how we address the two flaws mentioned in Sec. III-A. For flaw 1), object regions that are not activated in one sample by the classification objective can be re-activated by other samples, through searching prototypes with high confidence value O' and high similarity value s. Similar prototypes can be found with sufficient diversity of the training set and can re-activate via feature enhancement discussed in the following section. For flaw 2), if a prototype p_i belonging to the background is wrongly extracted, although its confidence scores O_i' is high, it is difficult to find regions that are highly similar to it on other training images due to the diversity assumption. Thus, its information propagation is limited.

Feature enhancement by re-weighting the feature maps. After we obtain the foreground probability maps FM, we can locate the object more accurately and comprehensively. We extract the feature maps of the last block $f_N(x,y)$ by multiplying with its confidence map FM (note that there are two FM and each corresponds to one of the image pairs). As the naive FM is sparse, we utilize a Gauss Filter [69] to smooth FM. In particular, we re-weight the original feature representations in the following way:

$$G(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}, \tag{9}$$

$$\mathbf{f_N}(\mathbf{x},\mathbf{y})^{'} = G(FM_{(x,y)}) \cdot \mathbf{f_N}(\mathbf{x},\mathbf{y}), \tag{10}$$

where $f_N(x,y)^{'}$ denote the enhanced object activation maps.

E. Training Process

We choose a multi-label soft cross-entropy classification loss $\mathcal{L}_c$ to make sure that the activated regions on $f_N^{'}$ still possess discrimination power to the object categories indicated by the image-level labels. To constraint the $f_N^{'}$ from deviating too much from the original feature maps f_N, which may cause divergence, we add a self-supervised loss $\mathcal{L}_{self}$ to compare both maps.

$$\mathcal{L} = \mathcal{L}_c + \lambda \mathcal{L}_{self}, \tag{11}$$

$$\begin{aligned} \mathcal{L}_c = -\sum_{i=1}^{N} (u[i] * log(\frac{1}{1+exp(-v[i])}) \\ + (1-u[i]) * log(exp(\frac{-v[i]}{1+exp(-v[i])}))), \end{aligned} \tag{12}$$

$$\mathcal{L}_{self} = \sum_{i=1}^{HW} (\mathbf{f'_N} - \mathbf{f_N})^2. \tag{13}$$

Here v denotes the predicted probability of class i while $u[i]$ denotes the image level groundtruth of i_{th} class. f_N and f'_N denote the object activation maps and its enhanced ones. We balance the losses with λ.

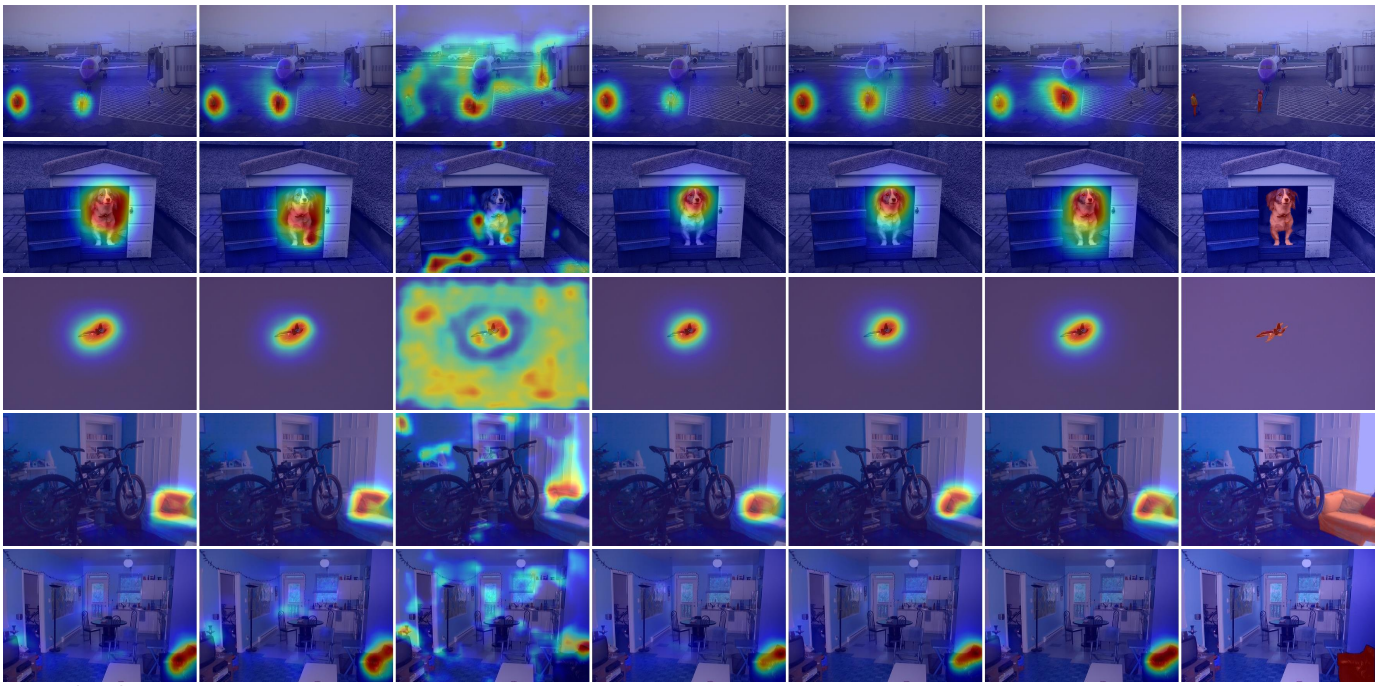

Fig. 4: The object activation maps trained with different numbers of the PASCAL VOC 2012 [20] training data. 1/2 denotes using only half the training data, 1/4 denotes using only quarter training data, and so on. Our baseline is IRN [4].

TABLE I: The performance of the synthesized segmentation labels evaluated on the PASCAL VOC 2012 and MS COCO training set. Base: base object activation maps, RPNet: our enhanced object activation maps. BR: Boundary Refine [4], CONTA [44].

Dataset	Base	Base + CONTA	Base + RPNet	Base + BR	Base + CONTA + BR	Base + RPNet + BR
VOC	48.3	48.8	**50.8**	66.0	67.9	**69.0**
COCO	27.4	28.7	**36.1**	33.9	35.2	**44.9**

TABLE II: Comparison of different training data volume on PASCAL VOC 2012. 1 denotes use all the VOC 2012 training data; 1/2 denotes use half of the dataset, and so on. When the training data decrease to 1/16, our RPNet can achieve comparable results, while the baseline's performance drops very quickly.

Data Volume	Baseline	Ours
1	48.3	52.8
1/2	48.0	51.5
1/4	46.6	50.5
1/8	41.1	49.6
1/16	18.2	46.9

TABLE III: Comparison different backbone. The results are evaluated on the PASCAL VOC 2012 training set.

Backbone	Enhanced	Base	Parameter	Top 1 error
resnet50	49.8	48.3	25.5M	23.8
resnet101	51.0	49.7	44.5M	22.6
xception	50.5	49.4	28.8M	21.2
wide_resnet50_2	49.7	48.4	68.9M	21.4
wide_resnet101_2	50.8	49.6	126.9M	21.1
resnext_50	50.8	49.1	25.0M	22.3
resnext_101	52.8	50.4	44.3M	20.6

IV. Experiment

A. Dataset and Evaluation Metric

Follow the experiment set-ups in previous works, we train and evaluate our RPNet on the PASCAL VOC 2012 [20] and MS COCO [25] dataset. The PASCAL VOC2012 training set is extended with images from [70] and finally gets 10582 training images, 1449 validation images, and 1456 testing images. MS-COCO [25] contains 81 classes, 80k training images, and 40k val images. We use image-level class labels only. We adopt the standard mean Intersection-over-Union(mIoU) as the evaluation metric for all of our experiments.

B. Implementation Details

All the backbones are pretrained on ImageNet [19] and we replace the stride of the last layer from 2 to 1 to maintain the feature map size. The batch size is set to 16. We apply the random crop, random scale, and random flip to the images for data augmentation during the training process. The learning rate is initially set to be 0.02 with a decrease rate of 0.9 at every iteration with polynomial decay [71]. We use SGD as

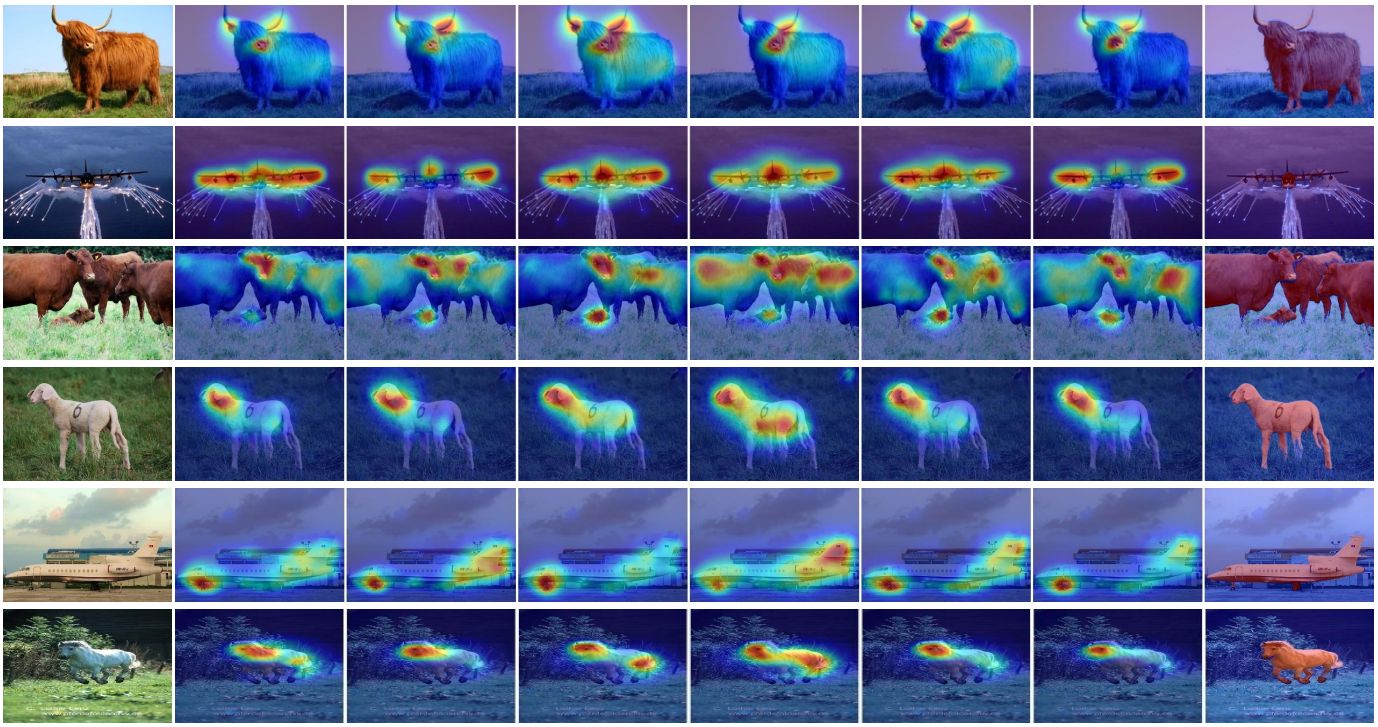

Fig. 5: Visualization examples that compare the object activation maps generated by the different backbones.

TABLE IV: Comparison different features combination.

block2	block3	block4	mIoU
✓			50
	✓		50.4
		✓	50.1
✓	✓		48.8
✓		✓	**50.8**
	✓	✓	50.1
✓	✓	✓	50.3

TABLE V: Comparison different probability to discard the object regions. The results are evaluated on the PASCAL VOC 2012 training set.

Probability	mIoU
0.1	50.1
0.3	50.2
0.5	50.8
0.7	50.5
0.9	50.3

the optimizer and train the network for 5 epochs. We set the probability β to 0.5 to discard the object activation regions as described in generate region prototypes, Sec. III-D , and set the σ to 3 as motioned in feature enhancement, equation 9, set the λ in equation 11 to be 10 and set the α to be 0.3 in equation 6. We adopt the boundary refinement method proposed in [4] and synthesize the pixel-level pseudo groundtruth to train a DeepLab-LargeFOV [72] with Resnet-50 backbone as our

TABLE VI: The performance of our RPNet with and without Gauss Filter. GF denotes Gauss Filter. The results are evaluated on the PASCAL VOC 2012 training set.

Ours w GF	Ours w/o GF
50.8	50.3

TABLE VII: The effectiveness of the relations between image pairs.

method	mIoU
baseline	48.3
prototype with one image	49.8
prototype with two image	50.8

semantic segmentation model. Finally, we follow the previous works [24], [73], [10], [6] to utilize denseCRF [74] as post-processing to further refine our predictions.

C. Ablative Analysis

In this section, if not specified, we employ Resnext-50 as the network backbone as default; all experiments are conducted on PASCAL VOC 2012 training set and evaluated with standard mIoU.

The Suitable Backbone. As is shown in Figure 5, we compare backbones and argue that the grouped convolution achieves a significant positive influence on the weakly supervised segmentation task: each channel group could learn a unique representation of the data [75] and the fused result

TABLE VIII: Comparison of the weakly supervised semantic segmentation methods. With the setting of without any additional supervision, our method outperforms all the previous methods on both validation set and test set.

Method	AS	Val	Test
FCN-MIL [13]	-	25.7	24.9
CCNN [47]	-	35.3	35.6
EM-Adapt [48]	-	38.2	39.6
DCSM [14]	-	44.1	45.1
BFBP [49]	-	46.6	48.0
SEC [10]	-	50.7	51.7
CBTS [11]	-	52.8	53.7
TPL [50]	-	53.1	53.8
MEFF [15]	-	-	55.6
PSA [5]	-	61.7	63.7
IRN [4]	-	63.5	64.8
SSDD. [6]	-	64.9	65.5
CIAN [23]	-	62.4	65.3
MBMNet [51]	-	62.4	65.3
MCI [52]	-	66.2	66.9
SEAM [12]	-	64.5	65.7
BENet [41]	-	65.7	66.6
CONTA [44]	-	65.3	66.1
RPNet (Resnet-50 w/o CRF)	-	**65.1**	**66.0**
RPNet (Resnet-50)	-	**66.4**	**67.2**
RPNet (Resnext-50 w/o CRF)	-	**65.7**	**66.7**
RPNet (Resnext-50)	-	**67.0**	**68.1**
RPNet (Resnext-101 w/o CRF)	-	**66.3**	**66.0**
RPNet (Resnext-101)	-	**68.0**	**68.2**

TABLE IX: Comparison of the weakly supervised semantic segmentation methods with more additional supervision, our method outperforms all the previous methods on both validation set and test set even we do not use any additional supervision. WV denote Web Video, S denote Saliency Mask, WI denote Web Image, IS denote Instance Image.

Method	AS	Val	Test
MIL-seg [53]	S+Img	42.0	40.6
MCNN [54]	WV	38.1	39.8
AFF [55]	S	54.3	55.5
STC [56]	S + WI	49.8	51.2
Oh et al. [57]	S	55.7	56.7
AE-PSL [42]	S	55.0	55.7
Hong et al. [21]	WV	58.1	58.7
WebS-i2 [7]	WI	53.4	55.3
DCSP [58]	S	60.8	61.9
GAIN [59]	S	55.3	56.8
MDC [60]	S	60.4	60.8
MCOF [61]	S	60.3	61.2
DSRG [62]	S	61.4	63.2
Shen et al. [8]	WI	63.0	63.9
SeeNet [24]	S	63.1	62.8
AISI [63]	IS	63.6	64.5
FickleNet [64]	S	64.9	65.3
DSRG+EP. [65]	S	61.5	62.7
Zeng et al. [66]	S	63.3	64.3
OAA+. [67]	S	65.2	66.4
RPNet (Resnet-50 w/o CRF)	-	**65.1**	**66.0**
RPNet (Resnet-50)	-	**66.4**	**67.2**
RPNet (Resnext-50 w/o CRF)	-	**65.7**	**66.7**
RPNet (Resnext-50)	-	**67.0**	**68.1**
RPNet (Resnext-101 w/o CRF)	-	**66.3**	**66.0**
RPNet (Resnext-101)	-	**68.0**	**68.2**

TABLE X: Comparison of the weakly supervised semantic segmentation methods with other methods on MS COCO dataset, our RPNet can outperforms all the previous methods.

Method	val
BFBP [68]	20.4
SEC [10]	22.4
SEAM [12]	31.9
IRNet [4]	32.6
CONTA [44]	33.4
RPNet	**38.6**

is more robust. We compare different backbone choices in Table III. We find that group convolution plays more important roles compared with the number of parameters and network structures.

Feature comparison. In Table IV, We compare our model variants that utilize different level features from our backbone encoder to generate the foreground probability maps. We experiment with the feature comparison with single block features and combining multiple block features. The best performance is achieved by combining block 2 and block 4. The possible explanation is that block2 corresponds to relatively low-level localization cues but lacks semantic cues while block4 is the opposite. This combination makes up for each other's shortcomings and gives play to their advantages, which leads to the final best performance. The block3 achieves the best performance with a single block is used; the possible explanation is that block3 contains both the low-level cues and the high-level cues consistent with our previous explanation.

The hyper-parameter choices. We investigate the choices of β, the possibility to discard a patch of the region during prototype generation. As shown in Table V, the best performance achieved when discarding the patches with the probability of 0.5.

The effectiveness of Gauss Filter. We utilize the Gauss Filter to smooth the naive sparse FM. We validate the effectiveness of the Gauss Filter in Table VI. The Gaussian filter can bring 0.5 mIoU improvement on the VOC 2012 training set.

The effectiveness of cross-image relations. To validate the effectiveness of our region prototype and the effectiveness of cross-image relations, we compare our network to a baseline model that does not employ our region prototype mechanism. As is shown in Table VII, when the region prototype vectors only from one image, our methods can bring 0.7 mIoU improvement over the baseline. While utilizing the cross-image relations, our RPNet can bring 1.7 mIoU improvement. More ablation studies can be found in supplemental materials.

D. Decrease the number of training samples

We investigate the robustness of the proposed method on datasets of reduced training samples. We reduce the number of training samples of each object category from PASCAL VOC 2012 and re-train our RPNet, and the baseline [4]. As shown in Table II, the performance baseline method [4] deteriorates

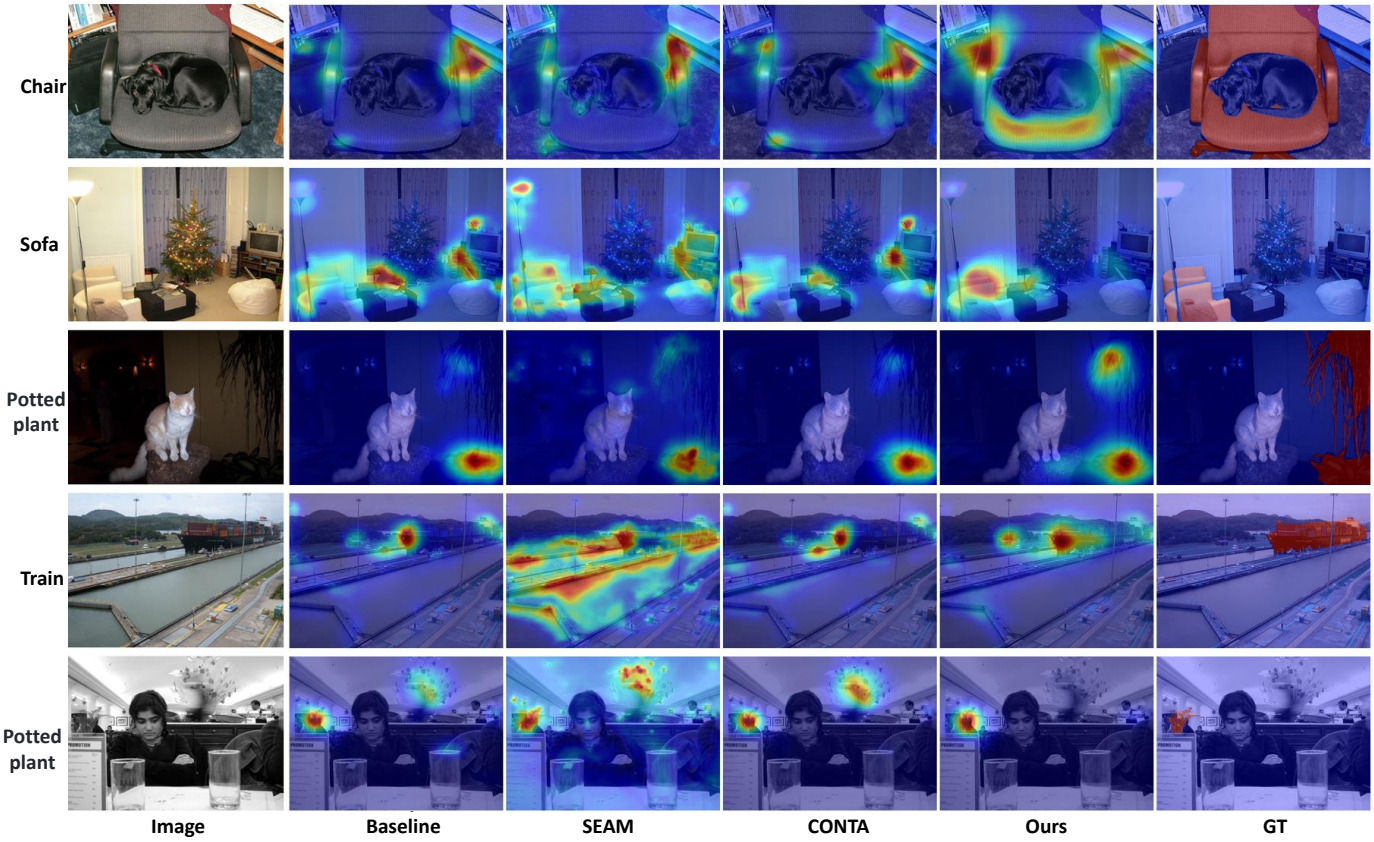

Fig. 6: Visualization examples that compare the object activation maps generated by different methods: SEAM [12], CONTA [44], and Ours: RPNet, GT denotes the ground truth. As shown in the figures, our RPNet can accurately predict the object localization and cover most object regions.

significantly as the number of samples is reduced to $1/16$ to its original number, but our proposed method's quality drops slowly. See Figure 4 for visual comparisons.

E. Analysis of Pseudo Labels

The performance of our synthesized segmentation labels is measured with standard mIoU on PASCAL VOC 2012 [20] and MS COCO [25] training data. As shown in Table I, with the same boundary refine method [4], the synthesized segmentation labels with enhanced object activation maps perform better than the base object activation maps and other state-of-the-art methods [44] on both datasets. With the same baseline(IRN) and Boundary Refine method, our RPNet performance 9.7 mIoU higher than the most current state-of-the-art method CONTA [44] on the large dataset MS COCO.

F. Comparison with State-of-the-art Results

For performance comparison, as shown in Table VIII and Table X, we achieve state-of-the-art performances on both datasets. Our method does not require any additional training information and even outperforms some methods with it, as in [58], [59], [62], [24], [63], [8].

Our RPNet outperforms the results of all previous methods with all types of backbone, such as Resnet-50, Resnext-50, and Resnext-101. We also verify our method with/without the denseCRF post-processing. It is observed that our method achieves the best performance regardless of denseCRF.

As shown in Table X, we evaluate our methods on the MS COCO dataset, which contains more classes and training images; this means the images have more diversity. Our RPNet achieves new state-of-the-art performances; to be noticed, our methods achieve more than 5 mIoU improvement over the other state-of-the-art methods. As shown in Table I, our RPNet can generate high-quality pseudos segmentation masks, which achieves 17.5 mIoU improvement over the baseline and 9.7 mIoU higher than the previous state-of-the-art methods.

We also show the qualitative examples on the PASCAL VOC 2012 in Figure 8 and MS COCO in Figure 7. It is particularly interesting to note that in the top left image of Figure 7, the ground truth misses an object(the handphone), but our prediction is able to segment it accurately. We also provide the detailed results on PASCAL VOC 2012 in Table XI.

V. Conclusion

We propose a novel framework to generate accurate pixel-level segmentation labels from better object activation maps with image-level labels only. Similar object parts across images are identified via region feature comparison. Object

Fig. 7: Our qualitative examples on the MS COCO.

TABLE XI: Detail results on the PASCAL VOC 2012 dataset. Our proposed method outperforms all previous methods on both val set and test set.

methods	bg	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	**mIoU**
PRNet (Resnet-50)	89.8	79.0	33.8	82.1	62.1	68.4	86.6	79.2	79.1	30.7	77.2	26.4	73.9	78.0	75.1	73.8	51.1	80.5	36.6	68.1	63.7	66.4
PRNet (Resnext-50)	89.7	69.4	32.8	82.2	67.2	71.1	87.7	78.7	80.3	32.8	79.1	34.7	76.5	80.9	74.2	74.8	53.2	80.0	42.5	65.2	55.0	67.0
PRNet (Resnext-101)	89.7	61.0	31.8	86.6	60.9	68.1	87.3	80.4	88.4	32.7	80.0	42.7	83.3	81.9	76.6	73.7	54.1	84.9	40.9	63.7	60.1	68.0

(a) The detail results on PASCAL VOC 2012 *val set*.

methods	bg	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	**mIoU**
PRNet (Resnet-50)	90.4	79.9	35.1	83.5	58.2	65.5	84.7	80.6	76.9	30.5	78.9	33.4	75.6	78.3	80.2	74.3	46.0	83.7	50.3	64.0	60.6	67.2
PRNet (Resnext-50)	90.3	77.5	33.5	83.1	59.7	68.6	86.8	81.1	79.0	31.5	81.1	40.5	79.9	81.9	81.9	74.9	49.3	85.7	53.3	57.8	52.9	68.1
PRNet (Resnext-101)	90.2	65.7	34.5	88.0	53.3	69.5	87.3	81.6	88.3	32.6	77.9	47.0	85.1	80.5	81.4	74.5	44.5	85.6	52.7	53.5	58.7	68.2

(b) The details results on PASCAL VOC 2012 *test set*.

confidence is propagated between regions to discover new object areas while background regions are suppressed. The performances on PASCAL VOC 2012 and MS COCO dataset validate our methods and achieve new state-of-the-art.

REFERENCES

[1] J. Dai, K. He, and J. Sun, "Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1635–1643.

[2] G. Papandreou, L.-C. Chen, K. Murphy, and A. Yuille, "Weakly-and semi-supervised learning of a dcnn for semantic image segmentation. arxiv, 2015," *arXiv preprint arXiv:1502.02734*.

[3] D. Lin, J. Dai, J. Jia, K. He, and J. Sun, "Scribblesup: Scribble-supervised convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3159–3167.

[4] J. Ahn, S. Cho, and S. Kwak, "Weakly supervised learning of instance segmentation with inter-pixel relations," in *CVPR*, 2019.

[5] J. Ahn and S. Kwak, "Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation," in *CVPR*, 2018.

[6] W. Shimoda and K. Yanai, "Self-supervised difference detection for weakly-supervised semantic segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 5208–5217.

[7] B. Jin, M. V. O. Segovia, and S. Susstrunk, "Webly supervised semantic segmentation," in *CVPR*, 2018.

[8] T. Shen, G. Lin, C. Shen, and R. Ian, "Bootstrapping the performance of webly supervised semantic segmentation," in *CVPR*, 2018.

[9] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu, "Learning fine-grained image similarity with deep ranking," in *CVPR*, 2014, pp. 1386–1393.

[10] A. Kolesnikov and C. H. Lampert, "Seed, expand and constrain: Three principles for weakly-supervised image segmentation," in *ECCV*, 2016.

[11] A. Roy and S. Todorovic, "Combining bottom-up, top-down, and smoothness cues for weakly supervised image segmentation," in *CVPR*, 2017.

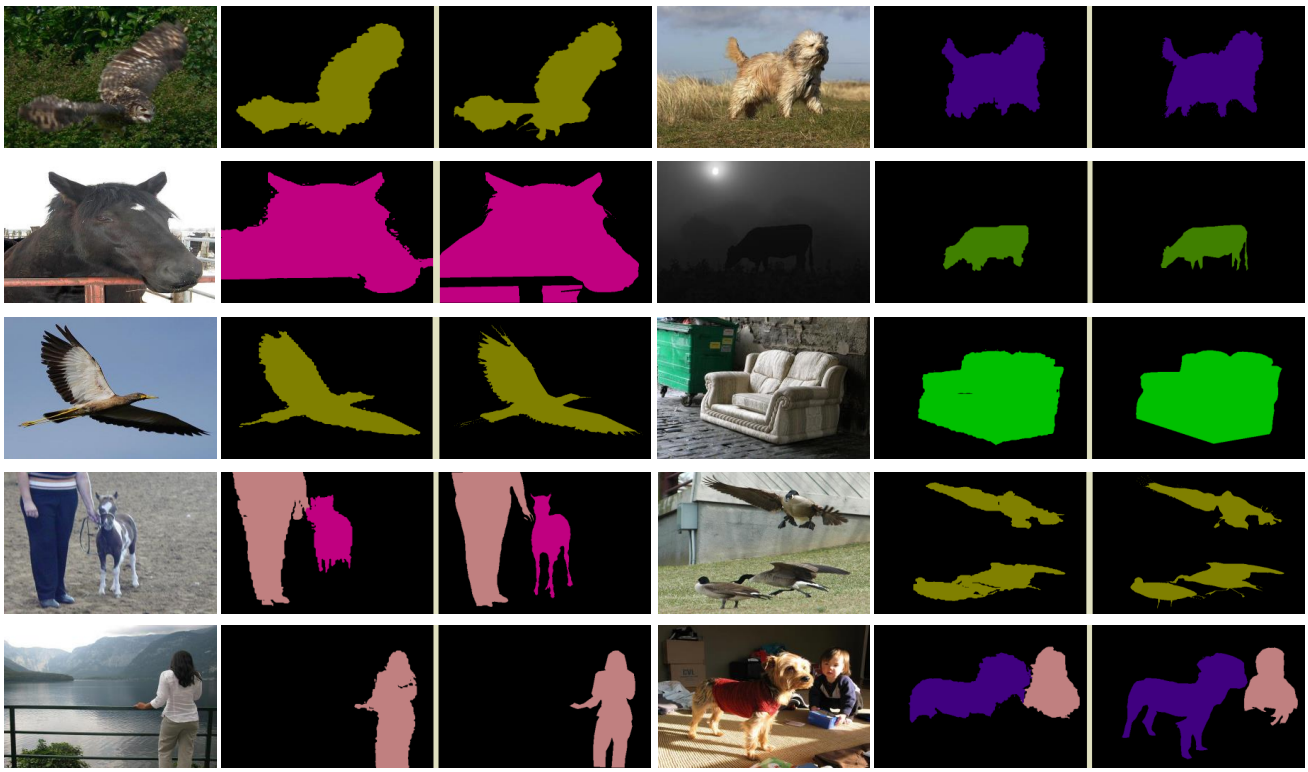

Fig. 8: Our qualitative examples on the PASCAL VOC 2012.

[12] Y. Wang, J. Zhang, M. Kan, S. Shan, and X. Chen, "Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12 275–12 284.

[13] D. Pathak, E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional multi-class multiple instance learning," in *ICLR*, 2015.

[14] W. Shimoda and K. Yanai, "Distinct class saliency maps for weakly supervised semantic segmentation," in *ECCV*, 2016.

[15] W. Ge, S. Yang, and Y. Yu, "Multi-evidence filtering and fusion for multi-label classification, object detection and semantic segmentation based on weakly supervised learning," in *CVPR*, 2018.

[16] J. Lee, E. Kim, S. Lee, J. Lee, and S. Yoon, "Frame-to-frame aggregation of active regions in web videos for weakly supervised semantic segmentation," in *ICCV*, 2019.

[17] T. Zhang, G. Lin, J. Cai, T. Shen, C. Shen, and A. C. Kot, "Decoupled spatial neural attention for weakly supervised semantic segmentation," *IEEE Transactions on Multimedia*, vol. 21, no. 11, pp. 2930–2941, 2019.

[18] A. Bearman, O. Russakovsky, V. Ferrari, and L. Fei-Fei, "What's the point: Semantic segmentation with point supervision," in *ECCV*, 2016.

[19] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*, 2009.

[20] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *International journal of computer vision*, vol. 111, no. 1, pp. 98–136, 2015.

[21] S. Hong, D. Yeo, S. Kwak, H. Lee, and B. Han, "Weakly supervised semantic segmentation using web-crawled videos," in *CVPR*, 2017.

[22] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.

[23] C. Song and J. Xiao, "Cian: Cross-image affinity net for weakly supervised semantic segmentation."

[24] Q. Hou, P.-T. Jiang, Y. Wei, and M.-M. Cheng, "Self-erasing network for integral object attention," in *NIPS*, 2018.

[25] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and D. P., "Microsoft coco: Common objects in context," in *ECCV*, 2014.

[26] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2018.

[27] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[28] H. Ding, X. Jiang, B. Shuai, A. Q. Liu, and G. Wang, "Semantic segmentation with context encoding and multi-path decoding," *IEEE Transactions on Image Processing*, vol. 29, pp. 3520–3533, 2020.

[29] S. Zhou, D. Nie, E. Adeli, J. Yin, J. Lian, and D. Shen, "High-resolution encoder–decoder networks for low-contrast medical image segmentation," *IEEE Transactions on Image Processing*, vol. 29, pp. 461–475, 2019.

[30] B. Kang, Y. Lee, and T. Q. Nguyen, "Depth-adaptive deep neural network for semantic segmentation," *IEEE Transactions on Multimedia*, vol. 20, no. 9, pp. 2478–2490, 2018.

[31] H. Shi, H. Li, F. Meng, Q. Wu, L. Xu, and K. N. Ngan, "Hierarchical parsing net: Semantic scene parsing from global scene to objects," *IEEE Transactions on Multimedia*, vol. 20, no. 10, pp. 2670–2682, 2018.

[32] W. Liu, C. Zhang, G. Lin, and F. Liu, "Crnet: Cross-reference networks for few-shot segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4165–4173.

[33] W. Liu, G. Lin, T. Zhang, and Z. Liu, "Guided co-segmentation network for fast video object segmentation," *IEEE Transactions on Circuits and Systems for Video Technology*, 2020.

[34] W. Liu, Z. Wu, H. Ding, F. Liu, J. Lin, and G. Lin, "Few-shot segmentation with global and local contrastive learning," *arXiv preprint arXiv:2108.05293*, 2021.

[35] Z. Wu, G. Lin, and J. Cai, "Keypoint based weakly supervised human parsing," *Image and Vision Computing*, vol. 91, p. 103801, 2019.

[36] Z. Wu, Q. Tao, G. Lin, and J. Cai, "Exploring bottom-up and top-down cues with attentive learning for webly supervised object detection,"

in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12 936–12 945.
[37] Z. Wu, X. Shi, J. Cai *et al.*, "Learning meta-class memory for few-shot semantic segmentation," *arXiv preprint arXiv:2108.02958*, 2021.
[38] Z. Wu, G. Lin, Q. Tao, and J. Cai, "M2e-try on net: Fashion from model to everyone," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 293–301.
[39] X. Shi, Z. Wu, G. Lin, J. Cai, and S. Joty, "Remember what you have drawn: Semantic image manipulation with memory," *arXiv preprint arXiv:2107.12579*, 2021.
[40] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *ICLR*, 2016.
[41] L. Chen, W. Wu, C. Fu, X. Han, and Y. Zhang, "Weakly supervised semantic segmentation with boundary exploration."
[42] Y. Wei, J. Feng, X. Liang, M.-M. Cheng, Y. Zhao, and S. Yan, "Object region mining with adversarial erasing: A simple classification to semantic segmentation approach," in *CVPR*, 2017.
[43] T. Zhang, G. Lin, W. Liu, J. Cai, and A. Kot, "Splitting vs. merging: Mining object regions with discrepancy and intersection loss for weakly supervised semantic segmentation."
[44] D. Zhang, H. Zhang, J. Tang, X. Hua, and Q. Sun, "Causal intervention for weakly-supervised semantic segmentation," *arXiv preprint arXiv:2009.12547*, 2020.
[45] X. Zhang, Y. Wei, and Y. Yang, "Inter-image communication for weakly supervised localization," 2020.
[46] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Advances in neural information processing systems*, 2017, pp. 4077–4087.
[47] D. Pathak, P. Krahenbuhl, and T. Darrell, "Constrained convolutional neural networks for weakly supervised segmentation," in *ICCV*, 2015.
[48] G. Papandreou, L.-C. Chen, K. Murphy, and A. L. Yuille, "Weakly-and semi-supervised learning of a dcnn for semantic image segmentation," in *ICCV*, 2015.
[49] F. Saleh, M. S. A. Akbarian, M. Salzmann, L. Petersson, S. Gould, and J. M. Alvares, "Built-in foreground/background prior for weakly-supervised semantic segmentation," in *ECCV*, 2016.
[50] D. Kim, D. Cho, D. Yoo, and I. S. Kweon, "Two-phase learning for weakly supervised object localization," in *ICCV*, 2017.
[51] W. Liu, C. Zhang, G. Lin, T.-Y. HUNG, and C. Miao, "Weakly supervised segmentation with maximum bipartite graph matching," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 2085–2094.
[52] G. Sun, W. Wang, J. Dai, and L. Van Gool, "Mining cross-image semantics for weakly supervised semantic segmentation," *arXiv preprint arXiv:2007.01947*, 2020.
[53] P. O. Pinheiro and R. Collobert, "From image-level to pixel-level labeling with convolutional networks," in *CVPR*, 2015.
[54] P. Tokmakov, K. Alahari, and C. Schmid, "Weakly-supervised semantic segmentation using motion cues," in *ECCV*, 2016.
[55] X. Qi, Z. Liu, J. Shi, H. Zhao, and J. Jia, "Augmented feedback in semantic segmentation under image level supervision," in *ECCV*, 2016.
[56] Y. Wei, X. Liang, Y. Chen, X. Shen, M.-M. Cheng, J. Feng, Y. Zhao, and S. Yan, "STC: A simple to complex framework for weakly-supervised semantic segmentation," in *IEEE Trans. on PAMI*, 2017.
[57] J. O. Seong, B. Rodrigo, K. Anna, A. Zeynep, and F. Mario, "Exploiting saliency for object segmentation from image level labels," in *CVPR*, 2017.
[58] A. Chaudhry, K. P. Dokania, and H. P. Torr, "Discovering class-specific pixels for weakly-supervised semantic segmentation," in *Proc. of British Machine Vision Conference*, 2017.
[59] K. Li, Z. Wu, K.-C. Peng, J. Ernest, and Y. Fu, "Tell me where to look: Guided attention inference network," in *CVPR*, 2018.
[60] Y. Wei, H. Xiao, H. Shi, Z. Jie, J. Feng, and T. S. Huang, "Revisiting dilated convolution: A simple approach for weakly- and semisupervised semantic segmentation," in *CVPR*, 2018.
[61] X. Wang, S. You, X. Li, and H. Ma, "Weakly-supervised semantic segmentation by iteratively mining common object features," in *CVPR*, 2018.
[62] Z. Huang, W. Xinggang, W. Jiasi, W. Liu, and W. Jingdong, "Weakly-supervised semantic segmentation network with deep seeded region growing," in *CVPR*, 2018.
[63] R. Fan, Q. Hou, M.-M. Cheng, G. Yu, R. R. Martin, and S.-M. Hu, "Associating inter-image salient instances for weakly supervised semantic segmentation," in *ECCV*, 2018.
[64] J. Lee, E. Kim, S. Lee, J. Lee, and S. Yoon, "Ficklenet: Weakly and semi-supervised semantic image segmentation," in *CVPR*, 2019.
[65] W. Wan, J. Chen, T. Li, Y. Huang, J. Tian, C. Yu, and Y. Xue, "Information entropy based feature pooling for convolutional neural networks," in *ICCV*, 2019.
[66] Y. Zeng, Y. Zhuge, H. Lu, and L. Zhang, "Joint learning of saliency detection and weakly supervised semantic segmentation," in *ICCV*, 2019.
[67] P.-T. Jiang, Q. Hou, Y. Cao, M.-M. Cheng, Y. Wei, and H.-K. Xiong, "Integral object mining via online attention accumulation," in *ICCV*, 2019.
[68] F. Saleh, M. S. Aliakbarian, M. Salzmann, L. Petersson, S. Gould, and J. M. Alvarez, "Built-in foreground/background prior for weakly-supervised semantic segmentation," in *European Conference on Computer Vision*. Springer, 2016, pp. 413–432.
[69] J.-M. Geusebroek, A. W. Smeulders, and J. Van De Weijer, "Fast anisotropic gauss filtering," *IEEE transactions on image processing*, vol. 12, no. 8, pp. 938–943, 2003.
[70] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, and J. Malik, "Semantic contours from inverse detectors," in *2011 International Conference on Computer Vision*. IEEE, 2011, pp. 991–998.
[71] W. Liu, A. Rabinovich, and A. C. Berg, "Parsenet: Looking wider to see better," *arXiv preprint arXiv:1506.04579*, 2015.
[72] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.
[73] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," *arXiv preprint arXiv:1412.7062*, 2014.
[74] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," in *Advances in neural information processing systems*, 2011, pp. 109–117.
[75] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds., vol. 25. Curran Associates, Inc., 2012, pp. 1097–1105.