Weakly-Supervised Semantic Segmentation Network with Deep Seeded Region Growing

Zilong Huang[1], Xinggang Wang[1]*, Jiasi Wang[1], Wenyu Liu[1], and Jingdong Wang[2]
[1]School of Electronic Information and Communications, Huazhong University of Science and Technology
[2]Microsoft Research Asia
{hzl,xgwang,wangjiasi,liuwy}@hust.edu.cn jingdw@microsoft.com

Abstract

This paper studies the problem of learning image semantic segmentation networks only using image-level labels as supervision, which is important since it can significantly reduce human annotation efforts. Recent state-of-the-art methods on this problem first infer the sparse and discriminative regions for each object class using a deep classification network, then train semantic a segmentation network using the discriminative regions as supervision. Inspired by the traditional image segmentation methods of seeded region growing, we propose to train a semantic segmentation network starting from the discriminative regions and progressively increase the pixel-level supervision using by seeded region growing. The seeded region growing module is integrated in a deep segmentation network and can benefit from deep features. Different from conventional deep networks which have fixed/static labels, the proposed weakly-supervised network generates new labels using the contextual information within an image. The proposed method significantly outperforms the weakly-supervised semantic segmentation methods using static labels, and obtains the state-of-the-art performance, which are 63.2% mIoU score on the PASCAL VOC 2012 test set and 26.0% mIoU score on the COCO dataset.

1. Introduction

Deep Convolutional Neural Networks (DCNN) have achieved great successes on the image semantic segmentation problem [5, 18] thanks to a large amount of fully-annotated images. However, collecting large-scale accurate pixel-level annotation is time-consuming and typically requires substantial financial investments. Unlabeled and weakly-labeled visual data, however, can be collected in large amounts in a relatively fast and cheap manner. Therefore, a promising direction in the computer vision research is to develop object recognition methods that can learn from unlabeled or weakly labeled images [14, 32].

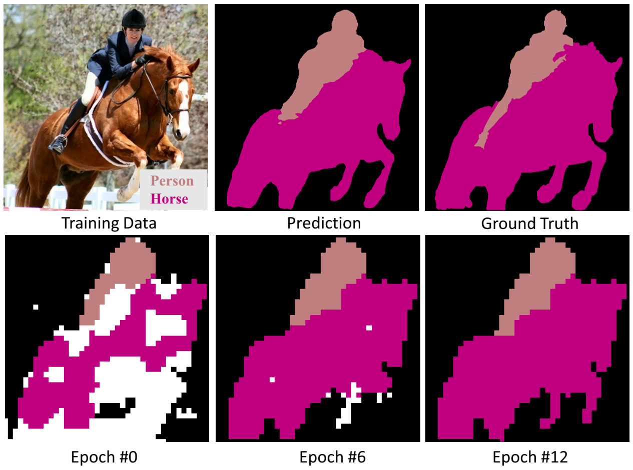

Figure 1. The top row orderly shows a training image with the image-level labels, the segmentation result of our proposed method only using image-level supervision, and the ground truth. Our segmentation result is very close to the ground truth annotated by human. The bottom row shows the dynamic supervision in several epochs during the training of the proposed weakly-supervised semantic segmentation network. (The black represents background and the white represents unlabeled/ignore pixels).

In this paper, we study the problem of learning semantic segmentation networks from weakly-labeled images. Among various settings of weak label, image-level annotation is one of the most economical and most efficient setting. In this context, every training image has its image class/category labels. It means objects belonging to the class labels appear in the image. However, the locations of the objects are unknown. We need to infer the pixel-level locations of the objects. Thus, the main problem in training weakly-supervised semantic segmentation networks is how to accurately assign image-level labels to their corresponding pixels.

To establish the desired pixel-label correspondence

*Corresponding author.

2575-7075/18 $31.00 © 2018 IEEE
DOI 10.1109/CVPR.2018.00733

in training, there is a very insightful research work. Kolesnikov et al. [14] employed an image classification network with classification activation maps (CAM) [37] method to select the most discriminative regions, and used the regions as pixel-level supervision for segmentation networks. Compared to the early weakly-supervised semantic segmentation methods [22, 20], the discriminative region based approach significantly improved the performance of this challenging task. However, in [14], the discriminative regions are small and sparse as shown in the epoch #0 image in Figure 1. In training, the supervision of the semantic segmentation network is fixed as the sparse discriminative regions. Thus, we name the learning strategy in [14] as "static supervision". The static supervision setting deviates from the requirement of semantic segmentation task that requires accurate and complete object regions for training segmentation models.

To address the issue, we propose to expand the discriminative regions to cover the whole objects during training segmentation networks. In practice, the pixels around the discriminative regions are always belonging to the same objects because semantic labels of the same object have spatial continuity. Our motivation is that, using the image labels enables to find small and sparse discriminative regions from the object of interest, termed as "seed cues", the neighboring pixels of seed cues with similar features (e.g. color, texture or deep features) could have the same labels as the seed cues. We utilize the classical Seeded Region Growing (SRG) method [1] to model this process for generating accurate and complete pixel-level labels. Here we can train semantic segmentation networks under supervision of the pixel-level labels. Different from [14, 19], the pixel-level labels are dynamic. The dynamic supervision is quite different from traditional network training using fixed supervision. In our case, we let the network generate new labels of the input training example, *i.e*., the training image. SRG is integrated into the deep segmentation network and can be optimized end-to-end and enjoys the deep features. We name the proposed method as "deep seeded region growing (DSRG)" for weakly-supervised semantic segmentation.

In practice, the seed cues localized by classification network is small but with high precision. It is a natural way to choose the seed cues as the seed points in SRG. Besides, to measure the similarity between the seed points and adjacent pixels for region growing, we make use of the segmentation map which is output of the segmentation network as features. Thus, SRG treats the seed cues as initial seed points; then the adjacent pixels in segmentation map with high probabilities on their corresponding categories take the same labels as the seed cues. This process is repeated until there are no pixels satisfying the above constraints. In the end, the output of DSRG is used as the supervision for training segmentation network. In the training phase, the supervision is used to form the loss function, termed as "seeding loss". In seeded regions, the loss is the same as full supervise loss function in [5]; the other positions are ignored by the seeding loss.

During training, the DSRG approach gradually enriches the supervision information of the segmentation network. As shown in Figure 1, the supervision in epoch #0 is actually the seed cues generated by classification model, the cues localize the head of person and the horse, which are the most discriminative regions in the image. With the increasing of epochs, the dynamic supervision gradually approaches the ground truth and cover the whole object content precisely. Meanwhile, the dynamic supervision ides the network to produce competitive segmentation result. To ensure the stability of training, DSRG always choose the original seed cues as initial seed points.

In the experiments, we demonstrate the effectiveness of our approach on the challenging PASCAL VOC 2012 Semantic Segmentation benchmark [8] and COCO, and show that we achieve the new state-of-the-art results. In addition, we provide an analysis of the DSRG approach by carrying out some ablation studies.

In summary, the main contributions of this paper are summarized below:

- In deep semantic segmentation network, we utilize the seeded region growing [1] mechanism, which enables the network safely generates new pixel-level labels for weakly-supervised semantic segmentation. Besides, the network can be optimized in an end-to-end manner and is easy to train.
- Our work obtains the state-of-the-art weakly-supervised semantic segmentation performance on the PASCAL VOC segmentation benchmark and COCO dataset. The mIoU of our method are 61.4% and 63.2% on pascal voc val set and test set respectively, which are better than many sophisticated systems and are getting closer to the fully supervised segmentation system [6] (67.6/70.3% mIoU on val/test set).

The rest of this paper is organized as follows. We first review related work in Section 2 and describe the architecture of our approach in Section 3. In Section 4, the detailed procedure to improve the quality of dynamic supervision is discussed and experimental results are analyzed. Section 5 presents our conclusion and future work.

2. Related work

The last years have seen a renewed interest on weakly-supervised visual learning. Various weakly-supervised methods have been proposed for learning to perform semantic segmentation with coarser annotations, such as image labels [20, 36], points [2], scribbles [16], and bounding boxes [7, 20] etc. In this work, we focus on using image labels as the main form of supervision, which is a simple supervision for training semantic segmentation models.

2.1. Pixel labeling from image level supervision

Pinheiro et al. [23] proposed a novel LSE pooling method which puts more weight on pixels which are important for classifying the image during training. Papandreou et al. [20] adopted an alternating training procedure based on the Expectation-Maximization algorithm to dynamically predict semantic foreground and background pixels. Qi et al. [24] proposed a unified framework that includes the semantic segmentation and object localization branches. [27] proposed a novel method to extract markedly more accurate masks from the pre-trained network itself. Wei et al. [35] presented a simple to complex learning method to gradually enhance the segmentation network. [29] proposed a method based on CNN-based class-specific saliency maps and fully-connected CRF. Roy et al. [26] presented a novel deep architecture which fuses three different cues toward semantic segmentation.

Recently, Kolesnikov et al. [14] proposed to localize seed cues according to classification networks for training segmentation network. However, [14] can only obtain small and sparse object-related seeds for supervision. To solve this problem, Oh et al. [19] proposed using a saliency model as additional information to exploit object extent. Wei et al. [33] used adversarial erasing manner to iteratively train multiple classification networks for expanding discriminative regions. Arslan et al. [4] also utilized adversarial erasing manner to allow the saliency detection network to discover new salient regions of object. Once true negative regions are generated, they have no chance to be correct them. In contrast, our proposed DSRG approach is very simple and convenient to start from the seed cues and progressively refine the pixel-level labels as the dynamic supervision in training phase.

Both [20] and the proposed method generate dynamic pixel-level labels to train semantic segmentation networks. However, there are several major improvements in this paper. Different from [20] where the latent pixel-level supervision is approximated by applying argmax function on biased segmentation maps, we instead propose to use the Seeded Region Growing to find accurate and reliable latent pixel-level supervision. With the help of the object seed cues, our DSRG training approach is robust to very noisy segmentation map in the beginning of training and generate pixel-level supervision with high accuracy all along.

2.2. Seeded Region Growing

The Seeded Region Growing (SRG) [1] is an unsupervised approach to segmentation that examines neighboring pixels of initial seed points and determines whether the pixel neighbors should be added to the region depending on a region similarity criterion. Two major concerns must be handled when performing a segmentation based on region growing: where to place the initial seeds in the image domain and which similarity criterion should be adopted to characterize the image regions. The most common way to select some seed pixels as seed based on simple hand-crafted criterion [28] (e.g. color, intensity, or texture). Meanwhile, the similarity criterion [3] is always defined on hand-crafted features. These settings result in over-segmentation and bad segmentation. In contrast, the DSRG utilizes seed cues generated by classification network as the initial seed to avoid wrong seed placement. Besides, We compute pixel similarity using deep learning features which have been proven to have high-level semantics. Thus, the DSRG can reduce over-segmentation and do not have the merge procedure of the traditional SRG.

3. Approach

In this section, we give the details of the proposed DSRG training approach for weakly-supervised semantic segmentation. At first, we will introduce how we generate seed cues from a deep classification network. Then, we will introduce a balanced seed loss function which uses seed cues as supervision to guide the weakly-supervised semantic segmentation network. At last, to address the problem that the seed cues are small and sparse, we propose the DSRG training.

3.1. Seed generation with classification network

We utilize a deep classification network to locate discriminative regions as seed cues under image-level supervision. Image-level labels do not explicitly provide any information about the position of semantic objects. But, recently, it has been shown that high-quality seeds indicating discriminative object regions can be obtained by learning a classification network under the supervision of image-level labels [30, 37]. The classification network is fully convolutional and the position of discriminative object regions are preserved in the deep layers of the network.

In our framework, we employ the CAMs [37] method for localizing the foreground classes. The procedures are briefly described as follows. We use a modified VGG-16 network [14] to initialize our classification network. In the network, global average pooling (GAP) is applied on conv7; the generated tensor is used as image representation and classified using a fully-connected layer; finally, the fully-connected classifier is applied to conv7 to generate a heatmap for each object class. Then the discriminative object regions are obtained by applying a hard threshold to the heatmap.

Besides of the seed cues in foreground, we also find seed cues in the background. For localizing background, we utilize the saliency detection technology from [12], and simply

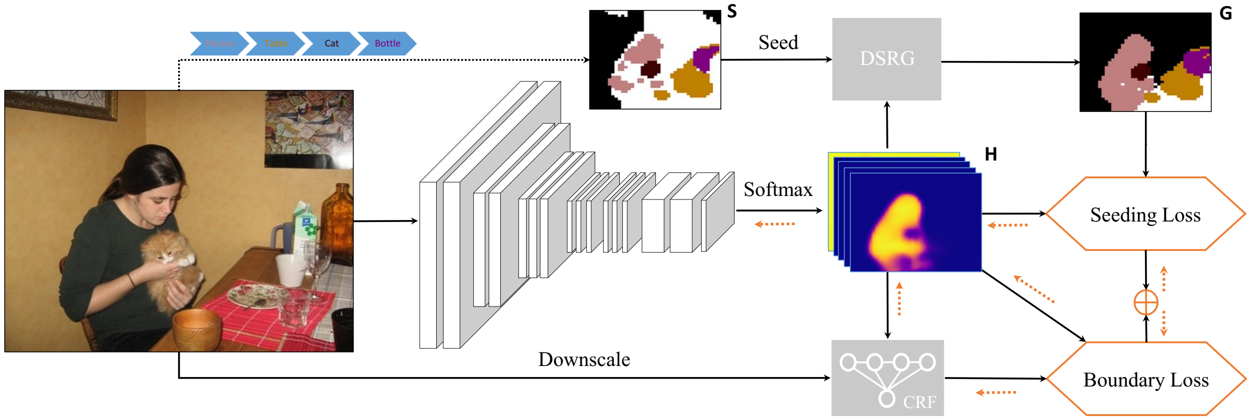

Figure 2. Overview of the proposed Deep Seeded Region Growing training approach. The Region Growing module takes the seed cues and segmentation map as input produces latent pixel-wise supervision which is more accurate and more complete than seed cues. Our method iterates between refining pixel-wise supervision and optimizing the parameters of a segmentation network.

select the regions in normalized saliency maps whose pixels are with low saliency values as background. The resulted seed cues from foreground and background are stacked together into a single channel segmentation mask.

3.2. Seeding loss

After obtaining the seed cues, we introduce how to train an image semantic segmentation network using the seed cues. The balanced seeding loss is proposed to encourage predictions of the segmentation network to match only seed cues given by the classification network while ignoring the rest of the pixels in the image. Considering the unbalanced distribution of the seed cues of foreground and background, the balanced seeding loss has two normalization coefficients for foreground and background, respectively, which is different from the seed loss in [14].

Let $\mathcal{C}$ be the set of classes that are present in the image (excluding background) and $\bar{\mathcal{C}}$ be the background. Suppose that S_c is a set of locations that are classified to class c. Then, the balanced seeding loss ℓ_{seed} is defined as follows:

$$\begin{aligned}\ell_{seed} = &-\frac{1}{\sum\limits_{c \in \mathcal{C}} |S_c|} \sum_{c \in \mathcal{C}} \sum_{u \in S_c} \log H_{u,c} \\ &-\frac{1}{\sum\limits_{c \in \bar{\mathcal{C}}} |S_c|} \sum_{c \in \bar{\mathcal{C}}} \sum_{u \in S_c} \log H_{u,c},\end{aligned} \tag{1}$$

in which $H_{u,c}$ denotes the probability of class c at position u of segmentation map H.

Besides, we use a boundary loss $\ell_{boundary}$ which proposed in [14] to encourage segmentation map to match up with object boundaries. Ultimately, the segmentation network are optimized by minimizing a loss function:

$$\ell = \ell_{seed} + \ell_{boundary}. \tag{2}$$

3.3. Deep seeded region growing

In the introduced seeding loss, we can find the seed cues are sparse. In practice, there are about 40% pixels have labels. During training, the labels are fixed following conventional setting of training deep networks. Our idea is to grow the seed cues to unlabeled pixels. Thus, we could have denser supervision to train better segmentation networks. The basis of seed cues growing is that in image there are small homogeneous regions in which the pixels should have the same label. The small homogeneous regions are usually used in low-level vision, such as generating superpixels [25]. To formulate the seed cues growing problem, here we refer to a classical algorithm, Seeded Region Growing (SRG) [1].

In SRG, some seed pixels are initially selected based on some simple hand-crafted criterion (e.g. color, intensity, or texture). Once the initial seeds are placed, the growth process seeks to obtain homogeneous image regions, i.e., it tries to segment the image into regions with the property that each connected component of a region contains exactly one of the initial seeds.

We propose to integrate SRG into deep segmentation networks for weakly-supervised semantic segmentation. The yield method is termed as "deep seeded region growing (DSRG)".

Once the initial seeds are initialized by classification network, the regions are then grown from these seed points to adjacent unlabeled points depending on a region similarity criterion. The similarity criterion defines whether a candidate pixel should be incorporated into a specific region or not. Now, the major concerns must be handled when performing learning a semantic segmentation network based on region growing: which similarity criterion should be adopted to characterize the image regions? In the following, we detail the strategies to handle the problem.

The similarity criteria P we make here is the simple probability threshold value of a pixel in segmentation map H generated by segmentation network.

$$P(H_{u,c}, \theta_c) = \begin{cases} \text{TRUE} & H_{u,c} \geq \theta_c \text{ and} \\ & c = \arg\max_{c'} H_{u,c'}, \\ \text{FALSE} & \text{otherwise.} \end{cases} \quad (3)$$

in which $H_{u,c}$ refers to the probability value of the pixel at position u that belongs to class c . And θ is the probability threshold value. In practice, we do not set different thresholds for different categories. The foreground categories share a same threshold θ_f and the background has another threshold θ_b. Traditional SRG usually has a phenomenon of over-segmentation since low-level image features is not robust to inter-class appearance of object. In DSRG, we compute pixel similarity using deep learning features which have been proven to have high-level semantics. Thus, the DSRG can reduce over-segmentation and do not have the merge procedure of the traditional SRG.

Now, we can take segmentation map H and seed cues S as inputs to perform region growing. DSRG is an iterative visiting process for each class. We denote the iterative visiting process of class c as $V_c, c \in [0, |C|]$, where $c = 0$ means the background class. In an iteration of V_c, we visit all the positions in S_c in a row-first manner. When visiting a pixel Q, we denote the set of unlabeled pixels in Q's 8-connectivity neighborhoods as R. For $R_u \in R$, its probability of being class c is denoted as $H_{u,c}$ as described above. Then R_u is classified based on P as follows:

```
if $P(H_{u,c}, \theta_c)$ then
    the pixel at $u$ is labeled as $c$;
else
    the pixel at $u$ keeps unlabeled state.
end if
```

After visiting all the positions, we append all the newly labeled pixels to S_c. Once S_c is changed, we will visit the updated S_c again. Otherwise, V_c stops. The termination criteria is different with classical SRG in which every pixel must have a label. Because it is difficult to tell the label of a pixel with a low confidence predicted by segmentation network. However, with increasing capability of segmentation network, the amount of unlabeled pixels decreases and the objects extent are covered with correct labels. Besides, to reduce the redundancy visits in V_c, we first compute connected components of regions that meet the requirement in Eqn (3), and then the connected components which consist the initial seed regions take the same label as the initial seed. These connected components are selected as new supervision for training segmentation network. We denote the $|C| + 1$ iterative visiting process as $DSRG(S, H)$, which means a region growing step. The final updated $S = [S_0, \cdots, S_C]$ is used as the supervision and applied to train segmentation network with seeding loss in Eqn (1). In Figure 2, the $DSRG(S, H)$ is plugged into the framework of the proposed segmentation network.

4. Experiments

4.1. Experimental setup

Dataset and Evaluation Metrics We evaluate the proposed approach on the PASCAL VOC 2012 segmentation benchmark dataset [8] and COCO dataset [17]. **PASCAL VOC:** It contains three parts: training (train, 1464 images), validation (val, 1449 images) and testing (test, 1456 images). Following the common practice [6, 33], we augment the training part by additional images from [9]. In our experiments, only image-level labels are utilized for training. We compare our method with other state-of-the-arts on both val and test sets. The standard intersection over union (IOU) criterion and pixel-wise accuracy are adopted for evaluation on PASCAL val dataset. The result on the test set is obtained by submitting the predicted results to the official PASCAL VOC evaluation server. **COCO:** its training set contains 80k samples with only image-level labels and it's val set contains 40k samples for evaluation. Performance is evaluated in terms of pixel IoU averaged on 81 categories. Experimental analysis of the proposed approach is conducted on the val set.

Training/Testing Settings

We adopt the slightly modified version of the 16-layer VGG network from [14] for the classification network and DeepLab-ASPP from [6] for the segmentation network. They are all initialized by the VGG-16 [31] pretrained on ImageNet. SGD with mini-batch is used for training classification and segmentation network. We use the momentum of 0.9 and a weight decay of 0.0005. The batch size is 20, the dropout rate is 0.5 and the weight decay parameter is 0.0005. The initial learning rate is 5e-4 and it is decreased by a factor of 10 every 2000 iterations.

For seed generation, those pixels belonging to top 20% of the largest value (a fraction suggested by [14, 33]) in the heatmap are considered as foreground object regions. We use saliency maps from [12] to produce the background localization cues. We adopt the normalized saliency value 0.06 as the threshold to obtain background localization cues (*i.e.* pixels whose saliency values are smaller than 0.06 are considered as background). For the similarity criteria in DSRG, we set θ_b and θ_f to 0.99 and 0.85, respectively. For CRF, we use the default values from the Koltun public implementation as parameters for the pairwise interactions.

In test phase, the learned segmentation network is ap-

Table 1. Comparison of weakly-supervised semantic segmentation methods on VOC 2012 val and test set

Method	Training	Val	Test
Supervision: Image-level Labels			
(* methods implicitly use pixel-level supervision)			
(† methods implicitly use box supervision)			
SN_B* [34]	10k	41.9	40.6
MIL-seg* [23]	700k	42.0	43.2
TransferNet* [10]	70k	52.1	51.2
AF-MCG* [24]	10k	54.3	55.5
GuidedSeg† [19]	20k	55.7	56.7
Supervision: Image-level Labels			
MIL-FCN [22]	10k	25.7	24.9
CCNN [21]	700k	35.3	35.6
MIL-bb [23]	700k	37.8	37.0
EM-Adapt [20]	10k	38.2	39.6
DCSM [29]	10k	44.1	45.1
BFBP [27]	10k	46.6	48.0
STC [35]	50k	49.8	51.2
SEC [14]	10k	50.7	51.7
AF-SS [24]	10k	52.6	52.7
Combining Cues [26]	10k	52.8	53.7
AE-PSL [33]	10k	55.0	55.7
DCSP [4]	10k	58.6	59.2
Supervision: Image-level Labels			
DSRG (VGG16)	10k	**59.0**	**60.4**[1]
DSRG (Resnet101)	10k	**61.4**	**63.2**[2]

plied to produce probability map for each testing image. Then, we upscale the predicted probability map to match the size of the input image, and then apply a fully-connected CRF [15] to refine the segmentation result.

Reproducibility. Our approach is implemented based on Caffe [11]. All networks are trained on a single NVIDIA GeForce GTX TITAN X GPU. The code is available at https://github.com/speedinghzl/DSRG.

4.2. Comparisons with state-of-the-arts

Results of other state-of-the-art weakly-supervised semantic segmentation solutions on PASCAL VOC validation and test dataset are summarized in Table 1. We provide these results for reference and emphasize that they should not be directly compared with our method. Because the methods were trained on different training sets or with different kinds of annotations, bounding boxes, spots and image-level labels. Among the approaches, CCNN [21], MIL-seg [23], STC [35], GuidedSeg [19], and TransferNet [10] use more images for training (700K, 700K, 50K, 20K and 70K, respectively). All the other methods are based on 10K training images and built on top of the VGG16 model.

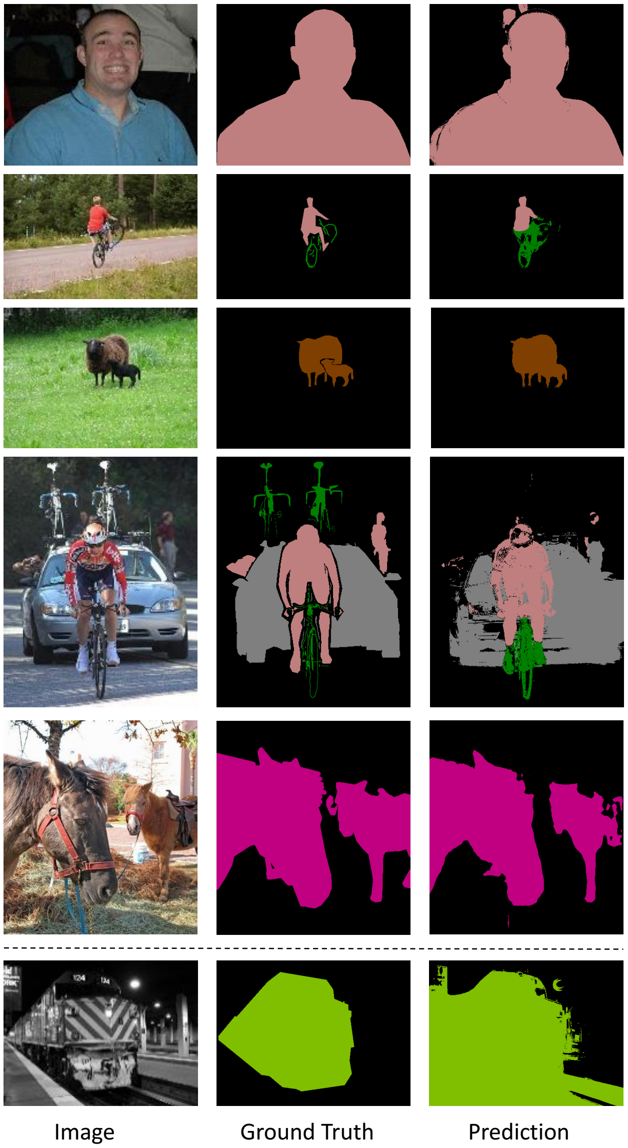

Figure 3. Qualitative segmentation results on the VOC 2012 val set. One failure case is shown in the last row.

The results show that our method substantially outperforms all the previous techniques using image-level labels for weak supervision. AE-PSL [33] and DCSP [4] achieve the best performance among the baselines. However, adver-

[1] http://host.robots.ox.ac.uk:8080/anonymous/ZZT4TI.html

[2] http://host.robots.ox.ac.uk:8080/anonymous/LWX93L.html

Table 2. Comparison of mIoU using different settings of our approach on VOC 2012 val set

Method	bkg	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	mIoU
baseline	82.5	67.5	23.2	65.7	29.7	47.5	71.8	66.8	76.7	23.3	51.7	26.2	69.7	54.2	63.2	57.2	33.7	64.5	33.5	48.7	46.1	52.5
+BSL	82.4	71.9	29.1	67.7	32.4	49.8	75.5	67.9	74.7	22.8	54.9	26.6	64.3	55.7	64.7	56.0	35.0	67.7	32.7	50.2	45.8	53.6
+DSRG	86.6	70.5	28.8	70.6	34.7	55.7	74.9	70.1	80.2	24.1	63.6	24.8	76.6	64.1	64.9	72.3	38.5	68.7	35.8	51.8	51.9	57.6
+Retrain	87.5	73.1	28.4	75.4	39.5	54.5	78.2	71.3	80.6	25.0	63.3	25.4	77.8	65.4	65.2	72.8	41.2	74.3	34.1	52.1	53.0	59.0

Table 3. Per-class IOU on COCO using image tags during training

Cat.	Class	SEC	BFBP	Ours	Cat.	Class	SEC	BFBP	Ours
BG	background	74.3	68.8	80.6	Kitchenware	wine glass	22.3	17.5	24.0
P	person	43.6	27.5			cup	17.9	5.6	20.4
Vehicle	bicycle	24.2	18.2	30.4		fork	1.8	0.5	0.0
	car	15.9	7.2	22.1		knife	1.4	1.0	5.0
	motorcycle	52.1	40.5	54.2		spoon	0.6	0.6	0.5
	airplane	36.6	32.0	45.2		bowl	12.5	13.3	18.8
	bus	37.7	39.2	38.7	Food	banana	43.6	44.9	46.4
	train	30.1	26.5	33.2		apple	23.6	18.9	24.3
	truck	24.1	17.5	25.9		sandwich	22.8	21.4	24.5
	boat	17.3	16.5	20.6		orange	44.3	35.0	41.2
Outdoor	traffic light	16.7	3.9	16.2		broccoli	36.8	27.0	35.7
	fire hydrant	55.9	33.1	60.4		carrot	6.7	16.0	15.3
	stop sign	48.4	28.4	51.0		hot dog	31.2	22.5	24.9
	parking meter	25.2	25.5	26.3		pizza	50.9	57.8	56.2
	bench	16.4	12.4	22.3		donut	32.8	36.2	34.2
Animal	bird	34.7	31.1	41.5		cake	12.0	17.0	6.9
	cat	57.2	52.8	62.2	Furniture	chair	7.8	8.2	9.7
	dog	45.2	44.1	55.6		couch	5.6	13.9	17.7
	horse	34.4	34.2	42.3		potted plant	6.2	7.4	14.3
	sheep	40.3	38.0	47.1		bed	23.4	29.8	32.4
	cow	41.4	42.1	49.3		dining table	0.0	2.0	3.8
	elephant	62.9	65.2	67.1		toilet	38.5	30.1	43.6
	bear	59.1	57.0	62.6	Electronics	tv	19.2	14.8	25.3
	zebra	59.8	65.0	63.2		laptop	20.1	19.9	21.1
	giraffe	48.8	55.6	54.3		mouse	3.5	0.4	0.9
Accessory	backpack	0.3	3.2	0.2		remote	17.5	9.9	20.6
	umbrella	26.0	28.1	35.3		keyboard	12.5	19.9	12.3
	handbag	0.5	1.1	0.7		cell phone	32.1	26.1	33.0
	tie	6.5	5.5	7.0	Appliance	microwave	8.2	9.8	11.2
	suitcase	16.7	21.3	23.4		oven	13.7	16.4	12.4
Sport	frisbee	12.3	5.6	13.0		toaster	0.0	0.0	0.0
	skis	1.6	1.0	1.5		sink	10.8	9.5	17.8
	snowboard	5.3	2.8	16.3		refrigerator	4.0	13.2	15.5
	sports ball	7.9	1.9	9.8	Indoor	book	0.4	7.5	12.3
	kite	9.1	10.3	17.4		clock	17.8	16.5	20.7
	baseball bat	1.0	1.7	4.8		vase	18.4	13.4	23.9
	baseball glove	0.6	0.5	1.2		scissors	16.5	12.2	17.3
	skateboard	7.1	6.6	14.4		teddy bear	47.0	41.0	46.3
	surfboard	7.7	3.3	13.5		hair dryer	0.0	0.0	0.0
	tennis racket	9.1	5.5	6.8		toothbrush	2.8	2.0	4.5
	bottle	13.2	9.6	22.3		**mean IOU**	**22.4**	**20.4**	**26.0**

sarial erasing is employed by AE-PSL to expand the seed cues for supervision, which needs to iteratively train multiple classification networks. DCSP also utilizes adversarial erasing manner to allow the saliency network to discover new salient regions of object. It does not require the retraining of the network after each erasing, but DCSP may introduce some true negative regions due to over erasing. In contrast, the proposed DSRG approach is very simple and convenient to refine supervision online and our method obtains better results than DCSP and AE-PSL. Compared with those methods only using image-level labels for supervision, the proposed DSRG(VGG16) method improves upon the best performance by over 1.2% on test set. It can be seen that our method achieves 60.4% mIoU on test set. Besides, Our DSRG (Resnet101) achieves 63.2% mIOU on test set.

4.3. Qualitative results

Fig. 3 shows some successful segmentation results. It shows our method can produce accurate segmentations even for complicated images and recover fine details of the boundary. One typical failure case is given in the bottom row of Fig. 3. This failure mode is that the model cannot pick out object regions from background precisely. As is typical for weakly-supervised systems, strongly cooccurring categories (such as train and rails, sculls and oars, snowbikes and snow) cannot be separated without finner-grained information [13].

4.4. Ablation studies

In order to further prove the effect of the different components, we conduct some ablation experiments with different settings of VGG16 based DSRG. In Table 2, the "baseline" denotes our implemented SEC [14], our result is much better than [14] (50.4 mAP without L_{expand}), due to the different background locating technology [12] and details. The "+BSL" denotes replacing the original seeding loss with the balanced seeding loss in Eqn (1); the "+DSRG" denotes adding DSRG training approach. We can observe that the weighted seeding loss improves the performance by 1.1% compared with baseline. And, DSRG improves further the performance by 4%, demonstrating the significant effectiveness of DSRG. It is most noticeable for animals and person, e.g. the improvement for segmenting dog/horse/cow/person is about 10%. Besides, we first employ the trained segmentation model of "+DSRG" to on all the training images. Then, the predicted segmentation masks are used as supervision for training the segmentation network for another round in a fully-supervised way. As shown in Table 2, the performance provided by this extra training (denoted as "+Retrain") is further improved from 57.6% to 59.0%. We do not observe further performance gain by performing additional retrain steps.

In addition, we tried different values of θ_f and θ_b to find the best performing region growing strategy. The re-

Table 4. Performance on PASCAL VOC 2012 val dataset for different θ

θ_f θ_b	0.99	0.95	0.90	0.85	0.80
0.99	57.45	57.59	57.63	**57.69**	57.66
0.95	57.43	57.56	57.64	57.67	57.63
0.90	57.23	57.35	57.40	57.44	57.45

sults are shown for different values of θ in Tab 4. The results show that our method is robust to the region growing thresholds θ. To explore the effect of only performing region growing for foreground or background object, we set $\theta_b = \infty, \theta_f = 0.85$ for only conducting region growing for foreground object, the performance on PASCAL VOC val dataset is 55.9% mIoU. When $\theta_b = 0.99, \theta_f = \infty$, the performance is 54.3% mIoU. The results show that only conducting region growing for foreground object or background object is also improve the performance. However, it can achieve best performance when simultaneously conducting region growing for foreground object and background object.

4.5. The quality improvement of dynamic supervision over epochs

In this section the qualities of the new pixel labels as dynamic supervision, obtained from DSRG, at each epoch, are evaluated. Compared with ground truths that are annotated by human, we could use the mean accuracy, mean recall and IoU to measure the quality of the supervision refined by our approach. In Fig. 4, the supervision that generated by classification network has somewhat high precision(62.6%), low recall(32.1%) and low IoU(30.0%). With the increasing of epochs, the precision of seed remains a high value, and the recall and IoU get significant improvements. At epoch #12, the mean precision, mean recall and mean IoU are 63.9%, 65.4%, and 57.1%, respectively. It demonstrates that DSRG can find the object extent and improve the quality of supervision, which explains why the proposed DSRG training procedure works excellently on the weakly supervised semantic segmentation task. Additional examples in the supplementary materials shows the gradually refining supervision starting from seed cues during training.

4.6. COCO results

To further demonstrate the generality of our method, we conducted a set of experiments on COCO. Unlike in PASCAL VOC, the majority of COCO samples were collected from non-iconic images in a complex natural context. We provide the per-class IoU of SEC [14], BFBP [27] and our approach in Table 3. Our VGG16 based DSRG obtains remarkable better results, especially in Person, Animal, Vehicle etc, but performs poorly on small ones, such as Indoor and Kitchenware. Altogether, our DSRG method improves upon the best performance by over 3.6% on val set. It can be seen that our method achieves 26.0% mIoU on val set. Meanwhile, compared with the performance of fully supervised method (40.98% mIoU), these results on COCO evidence that there is much space for progress in weakly-supervised semantic segmentation. Developing solutions that handle small objects could be an interesting direction for future research.

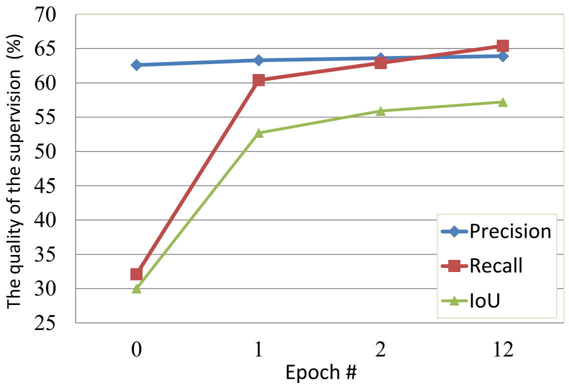

Figure 4. The quality of the dynamic supervision (%) with respect to the epochs.

5. Conclusion and future work

We have addressed the problem of training semantic segmentation networks only using image-level supervision. Image-level labels alone can provide high-quality seeds, or discriminative object regions, but inferring full object extents is a very difficult problem. We propose a DSRG training approach gradually improves the quality and extent object regions and itself is supervised the object regions. We demonstrate that our approach outperforms previous state-of-the-art methods under the same experimental conditions. We also clearly identify the effectiveness of region growing mechanism within the semantic segmentation network in the experiments. In future work, we will focus on designing more effective weakly-supervised strategies and improving seed quality.

Acknowledgements

We really appreciate the enormous help from Lichao Huang at Horizon Robotics and Chunyu Wang at MSRA. This work was partly supported by NSFC (No.61733007, No.61503145) and the fund of HUST-Horizon Computer Vision Research Center. Xinggang Wang was sponsored by CCF-Tencent Open Research Fund, the Program for HUST Academic Frontier Youth Team, and Young Elite Sponsorship Program by CAST, No. YESS 20150077.

References

[1] R. Adams and L. Bischof. Seeded region growing. *IEEE TPAMI*, 16(6):641–647, 1994. 2, 3, 4

[2] A. Bearman, O. Russakovsky, V. Ferrari, and L. Fei-Fei. Whats the point: Semantic segmentation with point supervision. In *Proc. ECCV*, pages 549–565. Springer, 2016. 2

[3] V. Borges, M. C. F. de Oliveira, T. Silva, A. Vieira, and B. Hamann. Region growing for segmenting green microalgae images. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2016. 3

[4] A. Chaudhry, P. K. Dokania, and P. H. Torr. Discovering class-specific pixels for weakly-supervised semantic segmentation. *arXiv preprint arXiv:1707.05821*, 2017. 3, 6

[5] 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. 1, 2

[6] 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. *arXiv preprint arXiv:1606.00915*, 2016. 2, 5

[7] J. Dai, K. He, and J. Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *Proc. ICCV*, pages 1635–1643, 2015. 3

[8] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *IJCV*, 111(1):98–136, 2015. 2, 5

[9] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, and J. Malik. Semantic contours from inverse detectors. In *Proc. ICCV*, pages 991–998. IEEE, 2011. 5

[10] S. Hong, J. Oh, H. Lee, and B. Han. Learning transferrable knowledge for semantic segmentation with deep convolutional neural network. In *Proc. CVPR*, pages 3204–3212, 2016. 6

[11] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proc. ACM*, pages 675–678. ACM, 2014. 6

[12] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li. Salient object detection: A discriminative regional feature integration approach. In *Proc. CVPR*, pages 2083–2090, 2013. 3, 5, 7

[13] A. Kolesnikov and C. H. Lampert. Improving weakly-supervised object localization by micro-annotation. *arXiv preprint arXiv:1605.05538*, 2016. 7

[14] A. Kolesnikov and C. H. Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *Proc. ECCV*, pages 695–711. Springer, 2016. 1, 2, 3, 4, 5, 6, 7, 8

[15] V. Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. *Proc. NIPS*, 2(3):4, 2011. 6

[16] D. Lin, J. Dai, J. Jia, K. He, and J. Sun. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *Proc. CVPR*, pages 3159–3167, 2016. 2

[17] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 5

[18] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proc. CVPR*, pages 3431–3440, 2015. 1

[19] S. J. Oh, R. Benenson, A. Khoreva, Z. Akata, M. Fritz, and B. Schiele. Exploiting saliency for object segmentation from image level labels. *arXiv preprint arXiv:1701.08261*, 2017. 2, 3, 6

[20] G. Papandreou, L.-C. Chen, K. Murphy, and A. L. Yuille. Weakly-and semi-supervised learning of a dcnn for semantic image segmentation. *arXiv preprint arXiv:1502.02734*, 2015. 2, 3, 6

[21] D. Pathak, P. Krahenbuhl, and T. Darrell. Constrained convolutional neural networks for weakly supervised segmentation. In *Proceedings CVPR*, pages 1796–1804, 2015. 6

[22] D. Pathak, E. Shelhamer, J. Long, and T. Darrell. Fully convolutional multi-class multiple instance learning. *arXiv preprint arXiv:1412.7144*, 2014. 2, 6

[23] P. O. Pinheiro and R. Collobert. From image-level to pixel-level labeling with convolutional networks. In *Proc. CVPR*, pages 1713–1721, 2015. 3, 6

[24] X. Qi, Z. Liu, J. Shi, H. Zhao, and J. Jia. Augmented feedback in semantic segmentation under image level supervision. In *Proc. ECCV*, pages 90–105. Springer, 2016. 3, 6

[25] X. Ren and J. Malik. Learning a classification model for segmentation. In *null*, page 10. IEEE, 2003. 4

[26] A. Roy and S. Todorovic. Combining bottom-up, top-down, and smoothness cues for weakly supervised image segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3529–3538, 2017. 3, 6

[27] F. Saleh, M. S. A. Akbarian, M. Salzmann, L. Petersson, S. Gould, and J. M. Alvarez. Built-in foreground/background prior for weakly-supervised semantic segmentation. In *Proc. ECCV*, pages 413–432. Springer, 2016. 3, 6, 8

[28] F. Y. Shih and S. Cheng. Automatic seeded region growing for color image segmentation. *Image and vision computing*, 23(10):877–886, 2005. 3

[29] W. Shimoda and K. Yanai. Distinct class-specific saliency maps for weakly supervised semantic segmentation. In *Proc. ECCV*, pages 218–234. Springer, 2016. 3, 6

[30] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013. 3

[31] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5

[32] P. Tang, X. Wang, X. Bai, and W. Liu. Multiple instance detection network with online instance classifier refinement. In *Proc. CVPR*, 2017. 1

[33] 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. *arXiv preprint arXiv:1703.08448*, 2017. 3, 5, 6

[34] Y. Wei, X. Liang, Y. Chen, Z. Jie, Y. Xiao, Y. Zhao, and S. Yan. Learning to segment with image-level annotations. *Pattern Recognition*, 59:234–244, 2016. 6

[35] 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. *IEEE TPAMI*, 2016. 3, 6

[36] J. Xu, A. G. Schwing, and R. Urtasun. Learning to segment under various forms of weak supervision. In *Proc. CVPR*, pages 3781–3790, 2015. 2

[37] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *Proc. CVPR*, pages 2921–2929, 2016. 2, 3