Box-driven Class-wise Region Masking and Filling Rate Guided Loss for Weakly Supervised Semantic Segmentation

Chunfeng Song[1,2] Yan Huang[1,2] Wanli Ouyang[3] Liang Wang[1,2,4,5]
[1]Center for Research on Intelligent Perception and Computing (CRIPAC),
National Laboratory of Pattern Recognition (NLPR),
Institute of Automation, Chinese Academy of Sciences (CASIA)
[2]University of Chinese Academy of Sciences (UCAS)
[3]The University of Sydney, SenseTime Computer Vision Research Group, Australia
[4]Center for Excellence in Brain Science and Intelligence Technology (CEBSIT)
[5]Chinese Academy of Sciences - Artificial Intelligence Research (CAS-AIR)
{chunfeng.song, yhuang, wangliang}@nlpr.ia.ac.cn wanli.ouyang@sydney.edu.au

Abstract

Semantic segmentation has achieved huge progress via adopting deep Fully Convolutional Networks (FCN). However, the performance of FCN based models severely rely on the amounts of pixel-level annotations which are expensive and time-consuming. To address this problem, it is a good choice to learn to segment with weak supervision from bounding boxes. How to make full use of the class-level and region-level supervisions from bounding boxes is the critical challenge for the weakly supervised learning task. In this paper, we first introduce a box-driven class-wise masking model (BCM) to remove irrelevant regions of each class. Moreover, based on the pixel-level segment proposal generated from the bounding box supervision, we could calculate the mean filling rates of each class to serve as an important prior cue, then we propose a filling rate guided adaptive loss (FR-Loss) to help the model ignore the wrongly labeled pixels in proposals. Unlike previous methods directly training models with the fixed individual segment proposals, our method can adjust the model learning with global statistical information. Thus it can help reduce the negative impacts from wrongly labeled proposals. We evaluate the proposed method on the challenging PASCAL VOC 2012 benchmark and compare with other methods. Extensive experimental results show that the proposed method is effective and achieves the state-of-the-art results.

1. Introduction

Semantic image segmentation refers to classifying each pixel in an image. Recently, semantic segmentation has achieved a series of progress [27, 38, 43, 8, 25, 30, 48, 17], among which [27] is the first to introduce Fully Convolutional Networks (FCN) structure into segmentation field. Following this work, there are some improvements through redesigning or adjusting the FCN structures [44, 18, 5, 31, 47, 7]. However, these works are designed for fully super-

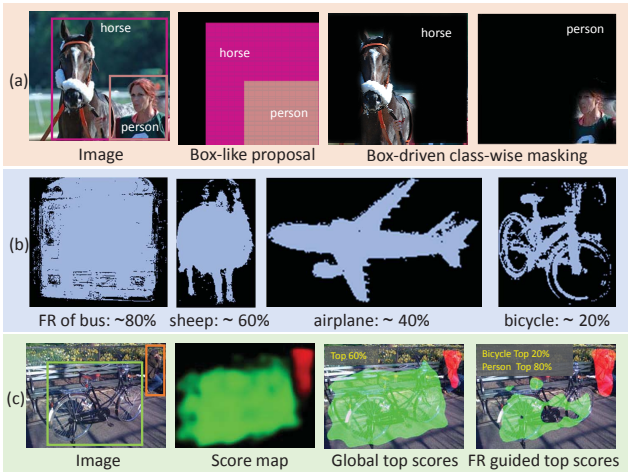

Figure 1. Weakly supervised segmentation with the box-level annotations. (a) The box-driven class-wise masking (BCM) model can learn specific masks for each class in region-level, and help remove the irrelevant regions of each class softly. (b) Based on the pixel-level segment proposals and the bounding boxes, we could calculate the mean pixel filling rates of each class, e.g., the sheep fills roughly 60% pixels of the box. (c) Via ranking the values of the score map, we can select the most confident locations for back propagation and ignore the weak ones. As shown in the picture, filling rate guided top scores selection is better than the global one.

vised mode, which has to be trained with large amounts of fully labeled data. Unlike other classic visual tasks such as classification and object detection, labeling semantic segmentation is rather expensive. For example, the cost of labeling a pixel-level segmentation annotation is about 15 times larger than labeling a bounding box, and 60 times than labeling an image class [26]. Considering bounding boxes also contain abundant semantic and objective information, a straightforward idea is to learn segmentation weakly with the bounding box supervision.

Recently, several weakly supervised segmentation methods [9, 29, 21, 2, 33] have been proposed to learn semantic masks with bounding box supervision. These methods mainly focus on generating high-quality pixel-level proposals. For example, in [29], the unsupervised dense CRF [22] was applied to eliminate the background within the bounding box. SDI [21] tried to produce segment proposals via combining MCG [32] and GarbCut [34] methods. BoxSup [9] updated the candidate masks generated by MCG in an iterative way. Then taking these enhanced segmentation proposals as pixel-level supervision, the deep FCN model can be trained for weakly supervised segmentation. Therefore, it is a core problem how to guide the FCN model to focus on the correct object regions and ignore the wrongly labeled regions from the segment proposals. Most previous approaches train the models with fixed proposals or simple iterative training. In this case, the gap between the ground-truth annotations and generated proposals limits their performance. We address this problem from two aspects.

First, considering that bounding boxes contain strong semantic and objective information, they should help us to remove the irrelevant regions and focus on the foreground regions. A straightforward idea is to learn a global mask to help remove the backgrounds in the images. However, the global mask can not learn multiple accurate shape templates for each class at the same time. To this end, we explore to adopt a box-driven class-wise masking (BCM) model to filter the feature maps of each class with boxes supervision, as shown in Figure 1 (a). The learned class-wise masks can provide obvious shape and location hints for each object, which is useful for the following segmentation learning.

Second, filling rate is a useful guidance for obtaining pseudo labels. It is well known that the score map in well trained model has different response values, indicating the confidence of prediction. A natural idea is to select the locations with the most active scores for backward learning, whereas ignore the less confident ones, as shown in Figure 1 (c). However, it is difficult to determine the threshold in a weakly supervised task, especially that different classes may need different thresholds. As shown in Figure 1 (b), different classes usually has different shapes, e.g., bus has 80% foreground pixels within its box while bicycle only fills 20% pixels of the box. This phenomenon inspires us to compute the mean filling rates of each class. Taking the pixel-level segment proposals generated through unsupervised methods as pseudo labels, we could calculate the mean pixel filling rates of each class. We find that the percentage of foreground pixels within the bounding box should be similar for the same class. Whereas the pixel filling ratios of two classes are usually different. Since the segment proposal for single sample is usually not accurate, the mean filling rate for samples of the same class can provide a more stable guidance. Rethinking the discussion above for the mean pixel filling rates, it will be a good choice to guide the top score selection with the filling rate. Based on this motivation, we propose a filling rate guided adaptive loss (FR-loss) to adjust the pseudo labels. Considering the situation that two objects from the same class may have different filling rates due to the shape and pose varieties, we try to refine the filling rates via clustering each class into several sub-classes.

Based on the analysis above, we propose the box-driven class-wise region masking (BCM) model and filling rate guided loss (FR-loss) for weakly supervised semantic segmentation. Firstly, we implement the BCM via segmentation guided learning with a box-like supervision. The proposed BCM can help remove the irrelevant regions of each class softly. It also provides an obvious hint of the foreground region, which could greatly contribute to the segmentation learning. Secondly, we calculate the mean filling rates of each class with the given bounding boxes and the generated pixel-level pseudo proposals. Thus we propose a filling rate loss to help select the most confident locations in the score map for back propagation and ignore the wrongly labeled pixels in proposals. With BCM and FR-loss working together, we could achieve the best performance with weak box supervision. We evaluate the proposed method on the challenging PASCAL VOC2012 dataset [12] and compare with previous methods. Extensive experimental results demonstrate that the our method is effective and achieves the state-of-the-art results. The performance of proposed method is even comparable with the fully supervised model. We summarize our contributions as follows:

- We introduce the box-driven class-wise masking (BCM) model to help remove the irrelevant regions of each class. It also provides an obvious hint of the foreground region, which could directly contribute to the segmentation learning.
- Filling rate guided adaptive loss (FR-Loss) is proposed to help select the most confident locations in the score map for back propagation and ignore the wrongly labeled pixels in the proposals.
- Extensive experiments on PASCAL VOC 2012 benchmark demonstrate that the proposed method is effective and achieves the state-of-the-art results.

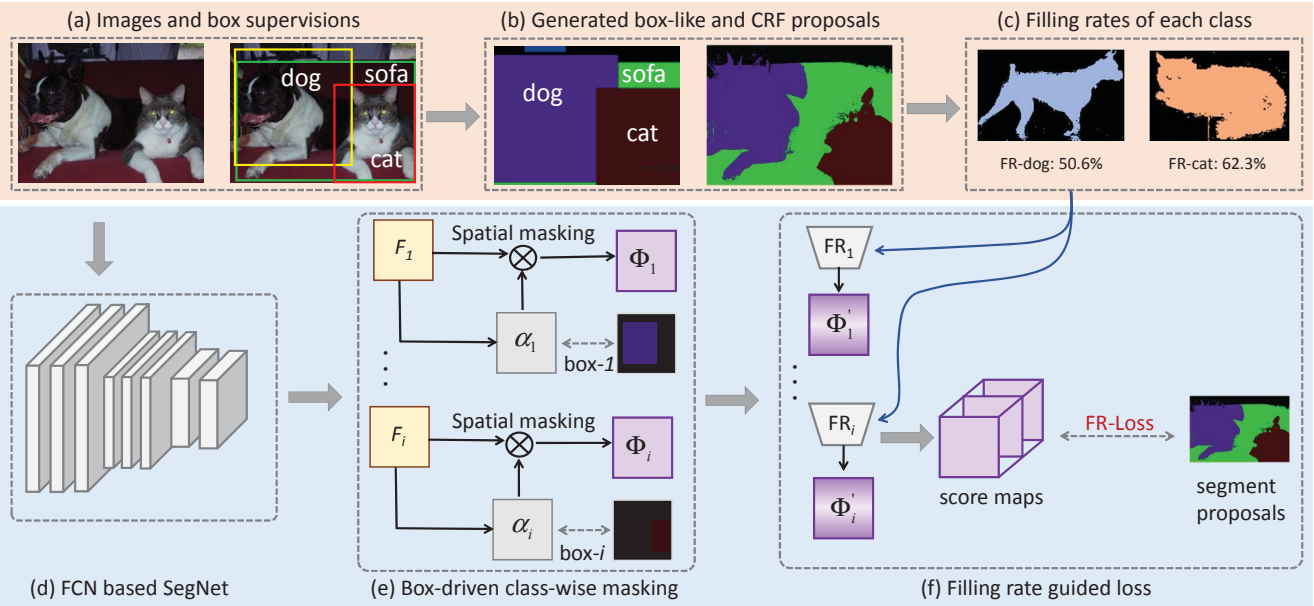

Figure 2. Pipeline of the proposed method. For a given image and its corresponding bounding boxes (a), we first generate the rectangle annotations (Box-like) and apply the unsupervised CRF [22] to generate segment proposals (b). We then calculate the mean filling rates of each class (c) with the CRF proposals and their corresponding boxes. With the images and segment proposals, we train the FCN based model (d), e.g., the DeepLab-LargeFOV network [5]. We add a box-driven class-wise masking (BCM) model (e) to generate class-aware masks via segmentation learning with box-like labels. The learned masks can implement spatial masking on the features of each class, separately. For each forward step, we rank the scores of each class in the prediction layer and adopt the filling rate guided loss (FR-loss) (f) to select the most confident locations for back propagation and ignore the weak ones. FR-loss could reduce the negative effects caused by the wrongly labeled pixels in the proposals.

2. Related Work

In this section, we briefly introduce the fully and weakly supervised semantic segmentation methods which are related to our work.

2.1. Fully Supervised Semantic Segmentation

Fully supervised semantic segmentation has achieved a series of progress [27, 38, 43, 8, 25, 30, 48, 17], among which [27] is the first to introduce the Fully Convolutional Neural Networks (FCN) structure into segmentation field. Following this work, a large number of improvements [44, 18, 5, 31, 47, 7, 14, 11, 37, 3, 10, 45] through redesigning or adjusting the network structures have been proposed. Chen et al. [5] introduce the atrous convolution for dense prediction and enlarge the receptive field of view. Zhen et al. [5] propose to adopt the dense CRF [22] with Gaussian pairwise potentials as a Recurrent Neural Network (RNN) to refine coarse outputs from a traditional CNN. Recently, an encoder-decoder based atrous separable convolution model [7] has achieved the state-of-the-art performance for fully supervised semantic image segregation.

2.2. Weakly Supervised Semantic Segmentation

Recently, a large number of weakly supervised methods explore to learn semantic segmentation with supervision of image labels [41, 19, 42, 1, 13], points [2], scribbles [39, 24, 28], and bounding boxes [9, 29, 21, 23]. The bounding boxes based methods are the most related works to this paper. BoxSup [9] introduces the recursive training procedure with the supervision of segment proposals. WSSL [29] proposes an expectation-maximization algorithm with segment proposals generated by the dense CRF [22]. Whereas SDI [21] tries to produce segment proposals via combining MCG [32] and GarbCut [34] methods. Li et al. [23] explore to segment the instance with both the bounding box supervision and the image tags. Different from these methods, we propose a box-driven class-wise masking (BCM) model to help remove the backgrounds before predicting the final segmentation. Unlike the global spatial attention model adopted in previous works [4, 6, 46, 15], the proposed BCM can learn specific attention maps for each class. To our knowledge, we are the first one to introduce the mean filling rate (FR) as a stable guidance through selecting the most confident locations in the score map for back propagation. The proposed FR-loss can adaptively select the re-

liable pixels and ignore the wrongly labeled pixels in the pseudo proposals.

3. Our Method

3.1. Overview

We present the proposed weakly supervised semantic segmentation framework with only bounding box supervision in this section. This framework can learn semantic masks from weakly box-level annotations via the box-driven class-wise masking (BCM) model and the filling rate guided loss (FR-Loss). In the following paragraph, we first describe the general pipeline, then introduce the details of each components.

There are mainly two steps for the proposed method, as shown in Figure 2. First, we generate the pixel-level proposals with the bounding box annotations and calculate the mean filling rates of each class. Then, we train the Fully Convolutional Network (FCN) based model with the proposed box-driven class-wise masking (BCM) model and filling rate guided loss (FR-loss).

Proposals Generating and Filling Rates Computing. The first step for weakly supervised semantic segmentation is to generate proper supervision labels from given bounding boxes, as shown in Figure 2 (b). The simplest yet widely used method is to convert the bounding boxes into rectangle segments directly, named as box-like proposals. Considering that the rectangle segments contain lots of wrongly labeled background regions within the bounding box, it is necessary to be further refined. There are several popular methods to generate high-quality segment proposals with bounding box labels, among which dense CRF [29], MCG [32] and GrabCut [34] are the mostly used approaches. For fair comparison with the baseline model [29], we choose the same unsupervised dense CRF as the default option to generate proposals. With the CRF proposals and their corresponding boxes, we can calculate the mean filling rates of each class, as shown in Figure 2 (c).

Model Training with BCM and FR-loss. As shown in Figure 2 (d), the backbone model in this paper is DeepLab-LargeFOV model [5]. Similar with the original FCN [27] training procedure, we also initialize this model with a VGG-16 model [36] pre-trained on ImageNet [35]. This backbone model is comparable with the ones used in the compared methods [9, 29, 21]. The FCN model takes the images as its inputs and the segment proposals as the supervision. To this end, the FCN model can be trained in an end-to-end manner. Note that in our case the quality of the supervision information in weakly supervised task is not guaranteed, so we add a box-driven class-wise masking (BCM) model to generate class-aware masks via segmentation learning with the box-like labels. The learned masks can implement spatial masking on the features of each class, separately. For each forward step, we rank the scores of each class in the prediction layer and adopt the filling rate guided loss (FR-loss) to select the most confident locations for back propagation and ignore the weak ones. FR-loss could reduce the negative effects caused by the wrongly labeled pixels in proposals. Details of them will be described in the next two sub-sections.

3.2. Box-driven Class-wise Masking

To remove the irrelevant regions in the feature maps, we need to learn specific masking maps for each class. Thus we design a box-driven class-wise masking (BCM) model to guide the learning of the segmentation model. We apply the masking on the FC-7 layer (note: implemented by convolution) of VGG-16 model [36] to mask the irrelevant regions. As shown in Figure 2 (e), the output features of FCN based SegNet are evenly sliced into N branches, corresponding to the N classes. For each branch, we add an binary attention model to produce a weights map for masking. To give a clear hint, we introduce the box-like mask to guide the attention map via adding a Mean Squared Error (MSE) loss on pixels of the attention map α_c and its corresponding mask M_c of class-c

$$L_{bcm(c)} = \sum_{h=1}^{H} \sum_{w=1}^{W} \left\| M_{c(h,w)} - \alpha_{c(h,w)} \right\|_2^2 \qquad (1)$$

where α_c has a size of (H, W). In the similar way, the N binary segmentation models can be trained separately. Then the N attention maps could carry out spatial-wise masking across their corresponding feature branches. We denote α_c and F_c as the learnt attention map and feature branch of class-c, respectively. Therefore, the weighted feature of class-c can be denoted as

$$\Phi_c = F_c \otimes \alpha_c \qquad (2)$$

where $\otimes$ means the spatial-wise masking operation. Then we combine the output features of N branches to produce the score map for final segmentation.

Unlike the global spatial attention model adopted in previous works [4, 6], the proposed class-wise masking model can learn specific attention maps for each class. It contributes to the segmentation models in three respects: 1) It can remove the irrelevant regions in the feature maps, such as the backgrounds. 2) It can learn N specific masking maps to fit each class, which may differ greatly from each other in shapes and sizes. 3) As the mask is learnt under the supervision of bounding box, thus it could provide a clear object hint for the segmentation learning.

3.3. Filling Rate Guided Adaptive Loss

Above box-driven class-wise masking model can guide the FCN to learn foreground features softly, we further ex-

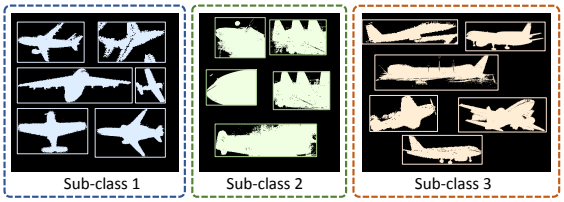

Figure 3. Examples of three sub-classes from airplane class. It is obvious that the middle sub-class has a larger filling rate than the right and left sub-classes. The mean filling rates of each sub-class could better represent different kinds of sample in one class.

plore to improve the segmentation learning in this subsection. Note that the wrongly labeled regions of the pixel-level proposals have negative effects on model training, recognizing the negative regions will be helpful. A possible solution is to ignore the pixels with small confident values in the score map, which may be the wrongly labeled pixels. In the weakly supervised mode, there are no guaranteed pixel-level annotations like the fully supervised mode, thus it is hard to determine how much percentage of pixels to be ignored. To address this problem, we introduce the filling rate guided adaptive loss (FR-loss). We intuitively find that the percentage of foreground pixels within the bounding box should be similar for the same class. Whereas the pixel filling ratios of two classes are usually different. Therefore, we first calculate the mean pixel filling rates of each class with pixel-level proposals and their corresponding boxes. For a given class-c, we denote the number of foreground pixels in the i-th proposal and box as $P_{proposal}(i)$ and $P_{box}(i)$, respectively. Then the mean filling rate of class-c can be defined as

$$FR_c = \frac{1}{N_c} \sum_{i=1}^{N_c} \frac{P_{proposal}(i)}{P_{box}(i)} \tag{3}$$

where N_c means the number of bounding boxes in class-c. Therefore, the mean filling rate of each class can be used to determine how much percent of the most confident pixels can be left for training or being ignored. In this way, the segmentation loss could be adjusted by the filling rates of each class. The FR-loss for one sample can be denoted as

$$L_{fr} = \sum_{c=1}^{N} \sum_{i=1}^{top(FR_c)} L_c(i) \tag{4}$$

where $L_c(i)$ means the loss of the i-th pixel with class-c, and the super-parameter top is determined by the mean filling rate of each class. This loss guides the score map to learn the most confident regions adaptively.

Refine the Filling Rates with Sub-class Clustering. Considering the situation that two objects from the same class may have different filling rates due to the shape and pose varieties, we try to refine the filling rates via k-meas clustering method [40] to classify each class into several sub-classes. As shown in Figure 3, we show the examples of three clustered sub-classes of airplane. Visually, three sub-classes are reasonable which can better represent three groups of boxes. Thus we take the mean filling rates of each sub-class to refine the FR-loss. In this situation, the FR-loss for one sample can be denoted as

$$L_{fr} = \sum_{c=1}^{N} \sum_{sc}^{3} \sum_{i=1}^{top(FR_{(c,sc)})} L_{(c,sc)}(i) \tag{5}$$

where $L_{(c,sc)}(i)$ means the loss of the i-th pixel with class-c and sub-class-sc. Note that $L_{(c,sc)}(i)$ is 0 when this pixel does not belong to this sub-class.

In retrospect, the class-wise masking model introduced in last subsection and the FR-loss can work together to guide the segmentation learning in a 'soft' manner, achieving comparable performance with the full-supervised model. The overall loss for one sample can be denoted as

$$L_{all} = L_{fr} + \lambda \cdot \sum_{c=1}^{N} L_{bcm(c)} \tag{6}$$

where λ is the hypermeter which is set to 0.01 in our experiments, N is the number of classes. We will evaluate the proposed methods in experiments.

Methods	Units	mIoU
Baseline [29]	-	60.6
Ours	CM	63.4
	BGM	64.9
	BCM	65.6
	Global-loss	64.1
	FR-loss	65.8
	FR-loss(Refine)	66.3
	BCM + FR-loss(Refine)	**66.8**

Table 1. Evaluate the effectivenesses of BCM and FR-loss on VOC2012 validation set. All models are based on the same Deeplab VGG16-LargeFOV backbones. The performance is evaluated in terms of mean IoU (%). CM: class-wise masking without box supervision, BGM: box-driven global masking, Global-loss: all boxes adopt the same global filling rate of 0.6.

4. Experiments

In the experiments, we first evaluate the effectiveness of our method on the Pascal VOC 2012 semantic segmentation dataset [12], then compare the proposed method with three state-of-the-art methods under weakly supervision and semi-supervision conditions separately.

4.1. Experimental Setup

Dataset. We evaluate the proposed framework on the widely used Pascal VOC2012 segmentation benchmark [12]. It contains 21 classes with pixel-level annotations. There are 1,464 images in the training set and 1,449 in the validation set, and the left 1,456 images are for testing. Following the same setting in the compared methods

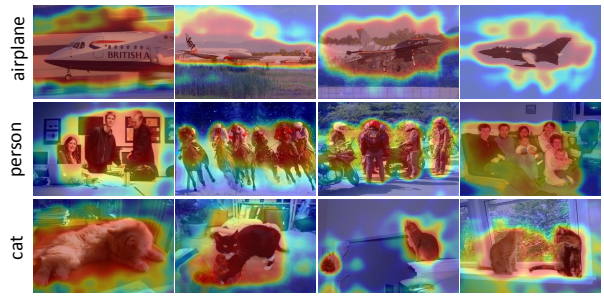

Figure 4. Visualization of the BCM learnt masking maps. It shows that most of the backgrounds are removed.

[9, 29, 21], we augment the training set with the data from SBD [16]. Consequently, there are 10,582 images in the training set and 1,449 images in the validation set. We train our model with the augmented training set and test it on the validation set to compare with other methods.

Implementation Details. We adopt the publicly released and widely used DeepLab-LargeFOV [5] model as the backbone network. It is based on a VGG-16 [36] network which has been pre-trained on ImageNet [35]. We train the proposed model under several different supervision settings. We first train the Deeplab-largeFOV model with the rectangle-box supervision. Further, we change the segments supervision into CRF-Box segments for finetuning, and regard it as the baseline model. Based on above model, we train the models with the proposed Box-driven Class-wise Masking (BCM) model and the Filling Rate guided Loss (FR-loss). We train the baseline model with roughly 20k iterations, and further fine-tune them with/without the BCM and FR-loss for 5k more iterations. In addition, we also evaluate the performance in the semi-supervised condition through adding 1,449 samples with ground-truth labels. The initial learning rates of the above models are 0.001 and decreased by 10 times after every 3k iterations, with a mini-batch size of 16/20 for the model with/without BCM. We take SGD as the default optimizer. For all the training phases, only flipping and cropping are adopted for data augmentation. With the well-trained FCN models, we can predict the semantic masks for the given images. Note the forward-passes of masking layers in BCM are parallel, the forward-passing time are very close to the baseline model, i.e., 42.7ms vs. 39.3ms per image. We also implement the dense-CRF [22] for post-processing on the masks. We adopt the same parameters of the dense-CRF with the compared work [29]. All experiments are implemented on a Nvidia TitanX GPU platform with the Caffe [20] framework.

Evaluation Metrics and Compared Methods. We adopt the "comp6" protocol to evaluate the performance. The accuracy is reported in terms of mean pixel Intersection-over-Union (mean IoU). We compare with three start-of-the-art methods (i.e., BoxSup [9], WSSL [29] and SDI [21]) on VOC 2012 dataset under both weakly supervised and semi-supervised conditions with bounding box annotations.

Modes	# GT	# Box	**Methods**	**mIoU**
Weak	-	10,582	BoxSup$_{Box}$ [9]	52.3
			WSSL$_{Box}$ [29]	52.5
			SDI$_{Box}$ [21]	61.2
			Ours$_{Box}$	54.9
			BoxSup$_{MCG}$ [9]	62.0
			WSSL$_{CRF}$ [29]	60.6
			SDI$_{M+G}$ [21]	65.7
			Ours$_{CRF}$	**66.8**
Semi	1,464	9,118	WSSL$_{Box}$ [29]	62.1
			BoxSup$_{MCG}$ [9]	63.5
			WSSL$_{CRF}$	65.1
			SDI$_{M+G}$ [21]	65.8
			Ours$_{CRF}$	**67.5**
Full	10,582	-	DeepLab-LargeFOV [5]	69.8

Table 2. Weakly and Semi-supervised results on VOC2012 validation set. With only 1/10 labeled segments, our method can achieve comparable performance with the fully supervised model. Box: directly using rectangle proposals, M+G: using the combined labels with both MCG and GrabCut.

Modes	# GT	# Box	**Methods**	**mIoU**
Weak	-	10,582	SDI [21]	69.4
			Ours	**70.2**
Semi	1,464	9,118	**Ours**	**71.6**
Full	10,582	-	DeepLab-ResNet-101 [5]	74.5

Table 3. Results of ResNet-101 backbone on VOC2012 validation set. Our method outperforms the compared SDI [21] method, achieving comparable performance with the fully supervised one.

4.2. Effectiveness of BCM and FR-Loss

We first evaluate the proposed framework with BCM and FR-loss, the results are shown in Table 1. Based on the Deeplab-LargeFOV model and CRF-box proposals, finetuning with the BCM model or the FR-loss can achieve 65.6% or 65.8% mean IoU accuracy, respectively. Both of them outperform the baseline model with obvious margins. When the BCM and FR-loss work together, we achieve 66.8% accuracy. The results show that the proposed BCM and FR-loss are effective and jointly combing the two modules can further enhance the performance. We also evaluate several variants of the proposed BCM and FR-loss. Experimental results show that the box-driven class-wise masking model performs better than the global one (BGM). We show the BCM learnt masks in Figure 4. Without the influence of the cluttered backgrounds, the segmentation learning could be more stable. It demonstrates that the class-wise attention model can guide the FCN model to learn more effective features, and the filling rate guided adaptive loss can help reduce the negative effects from the wrongly labeled proposals.

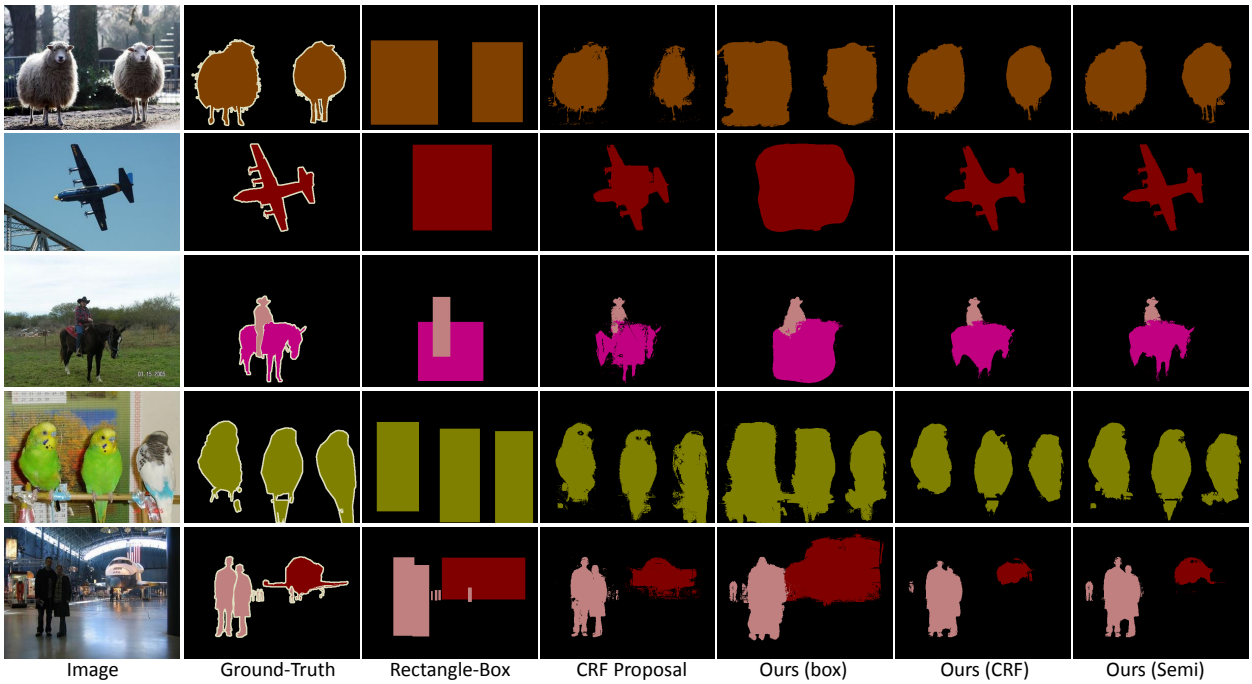

Figure 5. Examples of the segmentation results with proposed method. Original images are in the first column. The second column is the ground-truth segmentations. The 3-rd and 4-th columns are rectangle-box and CRF proposals. The following two columns show the results trained with rectangle-box and CRF proposals, respectively. The last column shows the results of semi-supervised model.

Methods	bkg	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	moto	person	plant	sheep	sofa	train	tv	**mean**
Weak(box)	78.3	37.4	20.6	46.6	44.9	64.5	80.7	68.1	59.8	32.5	65.7	58.4	61.6	51.2	53.2	60.5	47.5	60.0	49.3	64.2	49.4	**54.9**
Weak(CRF)	89.8	68.3	27.1	73.7	56.4	72.6	84.2	75.6	79.9	35.2	78.3	53.2	77.6	66.4	68.1	73.1	56.8	80.1	45.1	74.7	54.6	**66.8**
Semi	90.4	72.3	27.5	76.1	57.8	72.4	85.6	76.6	81.3	35.9	80.2	53.0	78.4	68.2	69.7	73.9	58.1	82.1	45.3	76.5	57.0	**67.5**

Table 4. Per class results of our method on VOC2012 validation set. The performance is evaluated in terms of mean IoU (%).

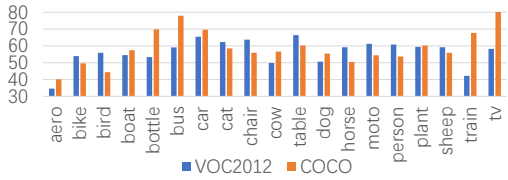

Figure 6. Filling rates of each class on VOC2012 and COCO. The filling rates are calculated with the generated pixel-level proposals. It is obvious that the filling rate can serve as an important cue for adjusting the pseudo labels.

4.3. Comparison with the State-of-the-art Methods

We compare with three state-of-the-art methods, i.e., BoxSup [9], WSSL [29] and SDI [21].

Results of Weakly-supervised Conditions. We first compare the results under the weakly supervised condition, as shown in Table 2. In this case, the only supervision label is the bounding box. We compare with BoxSup [9], WSSL [29] and SDI [21] from two aspects of view. Firstly, we compare the models trained with raw rectangle-box segments. The proposed method outperforms the BoxSup and WSSL, whereas SDI performs better which adopts an iterative training to update the segments from time to time. Secondly, we compare the models trained with the pre-processed segments. Our method outperforms all compared results and achieves an amazing performance with 66.8% mean IoU accuracy, which is very close to the full-supervised model. Note that our method adopts the same CRF-Box segments and the same base model with WSSL [29], whereas the performance of our method exceeds WSSL by roughly 6%. In addition, we compare the models trained with ResNet-101 backbone, as shown in Table 3. We achieve 70.2% mean IoU accuracy. The results demonstrate that the proposed method is effective for learning robust and accurate representations from bounding box annotations.

Results of Semi-supervised Conditions. We further compare with other methods in the semi-supervised task. In this task, 1,464 ground-truth labels are added for training. Although the amount of labeled samples is small which is only 1/10 of the training sets, they help improve the performance greatly. As shown in Table 2, the proposed method achieves 67.5% mean IoU accuracy, outperforming all the compared methods. With the extra 1/10 labeled segments, our model gets 0.7% improvement than its weakly

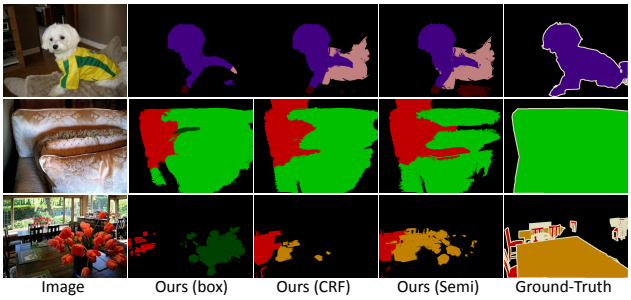

Figure 7. Failure examples of proposed method. Though our model achieves satisfying performances under both weakly and semi-supervised conditions, there are some frustrated results. For example, a dog wearing a cloth in the first image makes the model confused.

version. The results prove that our semi-supervised model can achieve comparable performance with the fully supervised model, showing the proposed BCM and FR-loss are still effective in semi-supervised mode.

4.4. Discussions

Above results have shown that the proposed method can learn better segmentations than the compared methods. To provide a comprehensive analysis, we report the per class results of proposed models, as shown in Table 4. We also calculate the per class FR of VOC2012 and COCO [26], as shown in Figure 6. It shows that the filling rates of VOC2012 and COCO are basically consistent, besides several classes, e.g., *train* and *tv*. It is obvious that among 21 classes, *airplane* and *sheep* are easy to segment, whereas *person* and *chair* are difficult. This result is consistent with the qualitative results shown in Figure 5 and 7. The generated CRF proposals can help the model learn pixel-level representations, achieving satisfied results. In addition, with the help of proposed BCM and FR-Loss, the model can reach a comparable performance with the fully supervised model. There are also some hard examples which bring great challenges to weakly supervised methods. As shown in Figure 7, it is very difficult for the model trained with limited and weakly labeled data to distinguish the classes in chaotic and complex scenes. This problem is worth to be deeply studied in the future work. Here, we will discuss the proposed methods separately.

Box-driven Class-wise Masking. With the class-level supervision, soft attention model based methods are widely adopted to guide the CNN model to learn better representations. Generally, the learned attention map usually contains object shape information. However, the global attention map can not learn multiple accurate shape templates for each class at the same time. In our method, the class-wise masking model can solve this problem. As shown in Figure 4, the learnt masks can remove the irrelevant regions and cluttered backgrounds to effectively contribute to the segmentation learning. In brief, BCM is helpful for box-driven weakly supervised segmentation through effective masking.

Filling Rate Guided Adaptive Loss. The FR-loss can guide the segmentation model to learn object masks in a soft manner, reducing the negative impacts from the wrongly labeled proposals. In this paper, we first directly set the mean filling rates of each class as the default value for determining the most confident locations. FR can be regarded as a kind of prior knowledge which could supervise the weakly learning procedure. Note that the filling rate of a class is independent of the others. The FR-loss is still effective when several classes have similar FR values and will not affect the performance. Considering that some samples may be greatly different from other samples though with the same class, the strategy of choosing top scores could be further improved. Thus we refine the filling rate via clustering each class into several sub-classes. It will be interesting to explore a better way to classify the sub-classes. We leave this problem as our future work.

5. Conclusion

In this paper, we have introduced a Box-driven Class-wise Masking (BCM) model to learn attention maps of each class. It can produce class-aware attentive maps for segmentation task learning, and provide an obvious hint whether this box or region contains a specific class. Moreover, based on the region-level segment proposals generated from the bounding boxes, we have proposed a Filling Rate guided adaptive loss (FR-loss) to help the model ignore the wrongly labeled pixels in proposals. FR-loss can adjust the model learning with global statistical information. The proposed BCM and FR-loss can work together to help reduce the negative impacts from wrongly labeled proposals. We evaluate the proposed method on the challenging PASCAL VOC 2012 benchmark and compare with other methods. Extensive experimental results show that the proposed method is effective and achieves the state-of-the-art results. In future, we will explore the jointly learning of the object detection and segmentation tasks to find more positive interactions between them.

Acknowledgement

This work is jointly supported by National Key Research and Development Program of China (2016YFB1001000), National Natural Science Foundation of China (61525306, 61633021, 61721004, 61420106015, 61806194), Capital Science and Technology Leading Talent Training Project (Z181100006318030), and Beijing Science and Technology Project (Z181100008918010). This work is also supported by grants from NVIDIA and the NVIDIA DGX-1 AI Supercomputer.

References

[1] Jiwoon Ahn and Suha Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *CVPR*, 2018. 3

[2] Amy Bearman, Olga Russakovsky, Vittorio Ferrari, and Li Fei-Fei. What's the point: Semantic segmentation with point supervision. In *ECCV*, 2016. 2, 3

[3] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. Hybrid task cascade for instance segmentation. *arXiv preprint arXiv:1901.07518*, 2019. 3

[4] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, and Tat-Seng Chua. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. *arXiv preprint arXiv:1611.05594*, 2016. 3, 4

[5] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. In *ICLR*, 2015. 1, 3, 4, 6

[6] Liang-Chieh Chen, Yi Yang, Jiang Wang, Wei Xu, and Alan L Yuille. Attention to scale: Scale-aware semantic image segmentation. In *CVPR*, 2016. 3, 4

[7] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. *arXiv preprint arXiv:1802.02611*, 2018. 1, 3

[8] Jifeng Dai, Kaiming He, Yi Li, Shaoqing Ren, and Jian Sun. Instance-sensitive fully convolutional networks. In *ECCV*, 2016. 1, 3

[9] Jifeng Dai, Kaiming He, and Jian Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *ICCV*, 2015. 2, 3, 4, 6, 7

[10] Xu Dan, Wanli Ouyang, Xiaogang Wang, and Nicu Sebe. Pad-net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing. In *CVPR*, 2018. 3

[11] Henghui Ding, Xudong Jiang, Bing Shuai, Ai Qun Liu, and Gang Wang. Context contrasted feature and gated multi-scale aggregation for scene segmentation. In *CVPR*, 2018. 3

[12] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *IJCV*, 2015. 2, 5

[13] Junsong Fan, Zhaoxiang Zhang, and Tieniu Tan. Cian: Cross-image affinity net for weakly supervised semantic segmentation. *arXiv preprint arXiv:1811.10842*, 2018. 3

[14] Zhang Hang, Kristin Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Ambrish Tyagi, and Amit Agrawal. Context encoding for semantic segmentation. In *CVPR*, 2018. 3

[15] Kota Hara, Ming-Yu Liu, Oncel Tuzel, and Amir-massoud Farahmand. Attentional network for visual object detection. *arXiv preprint arXiv:1702.01478*, 2017. 3

[16] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *ICCV*, 2011. 6

[17] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 1, 3

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1, 3

[19] Zilong Huang, Xinggang Wang, Jiasi Wang, Wenyu Liu, and Jingdong Wang. Weakly-supervised semantic segmentation network with deep seeded region growing. In *CVPR*, 2018. 3

[20] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM ICM*, 2014. 6

[21] Anna Khoreva, Rodrigo Benenson, Jan Hosang, Matthias Hein, and Bernt Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In *CVPR*, 2017. 2, 3, 4, 6, 7

[22] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *NeurIPS*, 2011. 2, 3, 6

[23] Qizhu Li, Anurag Arnab, and Philip HS Torr. Weakly-and semi-supervised panoptic segmentation. In *ECCV*, 2018. 3

[24] Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *CVPR*, 2016. 3

[25] Guosheng Lin, Chunhua Shen, Anton van den Hengel, and Ian Reid. Efficient piecewise training of deep structured models for semantic segmentation. In *CVPR*, 2016. 1, 3

[26] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 2, 8

[27] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 1, 3, 4

[28] Tang Meng, Federico Perazzi, Abdelaziz Djelouah, Ismail Ben Ayed, Christopher Schroers, and Yuri Boykov. On regularized losses for weakly-supervised cnn segmentation. In *ECCV*, 2018. 3

[29] George Papandreou, Liang-Chieh Chen, Kevin Murphy, and Alan L Yuille. Weakly-and semi-supervised learning of a dcnn for semantic image segmentation. In *ICCV*, 2015. 2, 3, 4, 5, 6, 7

[30] Pedro Pinheiro and Ronan Collobert. Recurrent convolutional neural networks for scene labeling. In *ICML*, 2014. 1, 3

[31] Pedro O Pinheiro, Ronan Collobert, and Piotr Dollar. Learning to segment object candidates. In *NeurIPS*, 2015. 1, 3

[32] Jordi Pont-Tuset, Pablo Arbelaez, Jonathan T Barron, Ferran Marques, and Jitendra Malik. Multiscale combinatorial grouping for image segmentation and object proposal generation. *IEEE TPAMI*, 2017. 2, 3, 4

[33] Carolina Redondo-Cabrera and Roberto J López-Sastre. Learning to exploit the prior network knowledge for weakly-supervised semantic segmentation. *arXiv preprint arXiv:1804.04882*, 2018. 2

[34] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. *ACM Transactions on Graphics*, 2004. 2, 3, 4

[35] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 2015. 4, 6

[36] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 4, 6

[37] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. Mask-guided contrastive attention model for person re-identification. In *CVPR*, 2018. 3

[38] Chunfeng Song, Yongzhen Huang, Zhenyu Wang, and Liang Wang. 1000fps human segmentation with deep convolutional neural networks. In *ACPR*, 2015. 1, 3

[39] Paul Vernaza and Manmohan Chandraker. Learning random-walk label propagation for weakly-supervised semantic segmentation. In *CVPR*, 2017. 3

[40] Kiri Wagstaff, Claire Cardie, Seth Rogers, and Stefan Schrödl. Constrained k-means clustering with background knowledge. In *ICML*, 2001. 5

[41] Xiang Wang, Shaodi You, Xi Li, and Huimin Ma. Weakly-supervised semantic segmentation by iteratively mining common object features. In *CVPR*, 2018. 3

[42] Yunchao Wei, Huaxin Xiao, Honghui Shi, Zequn Jie, Jiashi Feng, and Thomas S Huang. Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation. In *CVPR*, 2018. 3

[43] Zifeng Wu, Yongzhen Huang, Yinan Yu, Liang Wang, and Tieniu Tan. Early hierarchical contexts learned by convolutional networks for image segmentation. In *ICPR*, 2014. 1, 3

[44] Zifeng Wu, Chunhua Shen, and Anton van den Hengel. Wider or deeper: Revisiting the resnet model for visual recognition. *arXiv preprint arXiv:1611.10080*, 2016. 1, 3

[45] Dan Xu, Wanli Ouyang, Xavier Alamedapineda, Elisa Ricci, Xiaogang Wang, and Nicu Sebe. Learning deep structured multi-scale features using attention-gated crfs for contour prediction. In *NeurIPS*, 2017. 3

[46] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015. 3

[47] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015. 1, 3

[48] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip HS Torr. Conditional random fields as recurrent neural networks. In *ICCV*, 2015. 1, 3