Weakly-Supervised Segmentation by Combining CNN Feature Maps and Object Saliency Maps

Wataru Shimoda and Keiji Yanai
Department of Informatics, The University of Electro-Communications, Tokyo
1-5-1 Chofugaoka, Chofu-shi, Tokyo 182-8585 JAPAN
Email: {shimoda-k,yanai}@mm.inf.uec.ac.jp

Abstract—**In general, CNN based semantic segmentation methods assume pixel-wise annotation is available, which is costly to obtain in general. On the other hand, image-level annotations is much easier to obtain than pixel-level annotation. Then, in this work, we focus on weakly-supervised semantic segmentation which is known as task of using training data with only image-level annotations.**

In this paper, we propose a new CNN-based semantic segmentation method which uses both activation features calculated by feed-forwarding and object saliency maps obtained by back-propagation. As a CNN, we use the VGG-16 pre-trained with 1000-class ILSVRC datasets and fine-tuned it with multi-label training using only image-level labeled dataset. By the experiments, we show that the proposed method achieved state-of-the-art results with the PASCAL VOC 2012 dataset.

I. Introduction

Semantic image segmentation is a task to label object class labels to each of all the pixels in a given image, which is more challenging task than object classification and object detection. Semantic segmentation is expected to contribute detailed analysis of images in various practical tasks. Due to the recent advent of deep learning methods, convolutional neural network (CNN) based methods have outperformed most of the state-of-the-art in various kinds of image recognition tasks including semantic segmentation.

However, most of the CNN based semantic segmentation assume pixel-wise annotation is available, which is costly to obtain in general. On the other hand, collecting images with image-level annotation is relatively easier than pixel-level annotation, since there are many images attached with tags on the Web. Then, in this work, we focus on weakly-supervised semantic segmentation which requires not pixel-wise annotation as well as bounding box annotation but only image-level annotation.

In this work, we propose a new CNN-based semantic segmentation method which uses both feed-forwarding activations and object saliency maps obtained by back-propagation (BP). We integrate activation feature maps of convolutional layers of a CNN and BP-based object saliency maps [1] and use them to detect object regions.

Especially, we exploit and compare two kinds of feature maps based on the "Zoom-Out Features (ZOF)"[2] and "Fully Convolutional Network (FCN)"[3] for weakly supervised segmentation. Furthermore, we improve the method to estimate BP-based object saliency maps [1] for denser and clearer saliency maps by up-sampling saliency maps of the intermediate layers and aggregating them.

As a CNN, we use the VGG-16 model [4] pre-trained with 1000-class ILSVRC datasets and fine-tuned with multi-labeled training images in the PASCAL VOC dataset using only image-level labels.

To summarize our contributions in this paper, they are as follows:

- We propose a new method which uses both feed-forwarding feature maps and back-propagation based object saliency maps for weakly-supervised semantic segmentation.
- We show the effectiveness of the proposed method by the experiments with the Pascal VOC dataset, and achieved the state-of-the-art on the test dataset.

II. Related work

Recently, CNN-based semantic segmentation are being explored very actively, and the accuracy was much improved compared to the non-CNN-based conventional methods. In this section, first we describe full-supervised semantic segmentation, and next we explain weakly-supervised segmentation the objective of which is the same as our work.

A. CNN-based fully-supervised semantic segmentation

Girshick et al. [5] and Hariharan et al. [6] proposed an object segmentation method using region proposal and CNN-based image classification. Firstly, they generated 2000 region proposals at most by Selective Search [7], and secondly applied the trained CNN to each of the proposals. Finally they integrated all the CNN outputs and generated the final object regions. Although these methods outperformed the conventional methods greatly, they had a drawback that they required long processing time for CNN-based image classification of many region proposals. While Girshick et al. [5] and Hariharan et al. [6] took advantage of excellent ability of CNN on image classification task for semantic image segmentation in a relatively straightforward way, Long et al. [3] and Mostajabi et al. [2] proposed CNN-based semantic segmentation in a hierarchical way which achieved more robust and accurate segmentation.

Mostajabi et al. [2] proposed a method which associated up-sampled activation features of several intermediate layers with super-pixels, and treated them as local features, which are called "Zoom-Out Features (ZOF)". They achieved 69.6% accuracy on the Pascal VOC 2012 data set.

On the other hand, Long et al. [3] proposed a CNN-based segmentation method which integrated "deconvolution" and object heatmaps obtained by replacing all the full connection layers with 1×1 convolutional layers and providing a larger-size image than a usual 256×256 image. This modification replaced class score vectors with class score maps as outputs of the CNN, which expressed rough location of objects [8]. This idea was originally proposed by Sermanet et al. [9] and called as "Fully Convolutional Network (FCN)" or "sliding CNN", which played important roles to raise performance on CNN-based segmentation. By using larger-size images as input images, more detailed location information can be obtained in the intermediate layers as well as in the class score maps from the last layer. This can be used as unary priors of CRF [10], [11], [12].

In our work, we apply zoom-out features (ZOF) and fully convolutional network (FCN) both of which showed high performance on fully supervised segmentation task for weakly-supervised segmentation tasks. We use them to detect rough object location in the weakly-supervised setting.

B. *CNN-based weakly-supervised segmentation*

Simonyan et al. [1] showed that object segmentation without pixel-wise training data can be done by using back-propagation processing which is a method to train a CNN. To train a CNN, we optimize CNN parameters so as to minimize the loss between groundtruth values and output values. In the back-propagation process, derivatives of loss are propagated from the top layers to the lower layers. Springenberg et al. [13] also proposed a method for object localization by back-propagating the derivatives of a maximum loss value of the object detected. They achieved more accurate localization by limiting back-propagating values to positive values. However, in [1], [13], there are little difference in derivatives obtained from signals of each class. Therefore, we used derivatives for estimating background region and combining feed-forward activations.

Pedro et al. [14] achieved weakly-supervised segmentation by using multi-scale CNN proposed in [9]. They integrated the outputs which contained location information with log sum exponential, and limited object regions to the regions overlapped with object proposals.

Pathak et al. [15], [16] and Papandreou et al. [17] achieved weakly-supervised semantic segmentation by adapting CNN models from fully-supervised segmentation to weakly-supervised segmentation. In [15] they combined the output of the model proposed in [3] with global-max-pooling, and they enabled weakly-supervised training. In [16], they improved their method by adding some constraints. Papandreou et al. [17] trained the model proposed in [10] with EM algorithm.

As described above, recently in some works [15], [16], [17] the models for fully-supervised segmentation based on FCN [3] were adapted for weakly-supervised segmentation. However, "zoom-out features" proposed in [2] is not applied to weakly-supervised segmentation. Then, in this paper, we exploit zoom-out features for weakly-supervised semantic segmentation and compared a method based on FCN with it.

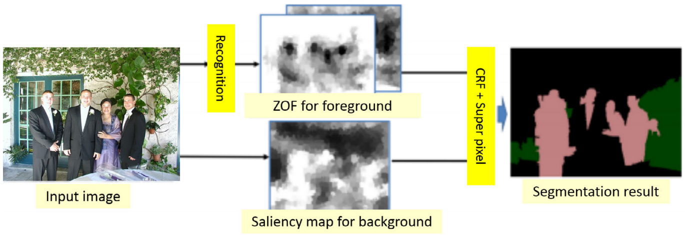

Fig. 1. The processing flow of the ZOF-base method.

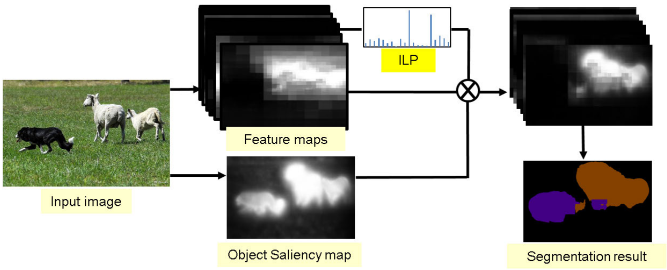

Fig. 2. The processing flow of the FCN-based method.

In addition, we use object saliency maps obtained by back-propagation [1] and aggregated intermediate derivatives by up-sampling to generate dense maps. We show saliency maps can improve inference based on feed-forward activations by treating as region likelihood of background class which is similar to smoothing priors on [17].

III. Method

In the proposed method, for weakly-supervised object region estimation, we use feed-forward feature maps of a CNN as well as back-propagation-based object saliency maps [1]. As the ways to use feature maps for region estimation, we exploit and compare two methods, "zoom-out features (ZOF)" [2] and "fully convolutional network (FCN)" [3], both of which achieved state-of-the-arts as fully-supervised segmentation methods. Figure 1 and 2 show the processing flow of two methods, ZOF-based method and FCN-based method, respectively.

A. *CNN model*

In this work, we use VGG-16 [4] as a basic CNN architecture. In our framework, we fine-tune a CNN with training images having no pixel-wise and bounding box information but image-level multi-label annotation. To carry out multi-label training of the CNN, we use Sigmoid cross entropy loss which is a standard loss function for multi-label annotation instead of soft-max loss. The Sigmoid cross entropy loss function is represented in the following equation:

$$\text{loss} = \sum_{k=1} [p_n \text{log}\, \hat{p_n} - (1 - p_n)\text{log}(1 - \hat{p_n})] \qquad (1)$$

where K is the number of classes, $p_n = \{0, 1\}$ which represents the existence of the corresponding class label, and $\hat{p_n}$ means the output of Sigmoid function of the class score $f_k(x)$ represented in the following equation:

$$\hat{p_n} = \frac{1}{1 + e^{-f_k(x)}} \tag{2}$$

B. Zoom-out features

Zoom-out features which has been proposed by Mostajabi et al [2] is a method which achieved the state-of-the-art for fully supervised semantic segmentation. In [2], they associated activation signals in feature maps with super-pixels of a given image, and obtained visual features of super-pixels by averaging signals over each of super-pixels. They also up-sampled all the feature maps so that their size became the same as a given image, and integrated up-sampled feature maps to estimate object locations more accurately.

In this paper, we apply "Zoom-Out Feature (ZOF)" [2] to weakly-supervised segmentation. Note that we use super-pixels as region representation in the same way as [2] for this ZOF-based method. In case of fully-supervised training, only ZOF inside the labeled object regions can be extracted as feature vectors for training, since pixel-wise annotation is available. However, in case of the weakly-supervised setting, correspondences between groundtruth labels and object regions are unknown. Then, we estimate correspondences using multiple instance learning (MIL) which is one of common methods to estimate regions corresponding to given labels, and adopt mi-SVM [18] as a method of multiple instance learning which uses SVM iteratively. Given a certain class, we regard images having the label of the target class as positive bags, and image having no label of the target class as negative bags. Positive bags contain more positive regions, while negative bags contain no positive regions. Because MIL can estimate positive regions, we can estimate positive super-pixels by using MIL.

In the proposed ZOF-based method, we integrate object super-pixels estimated by ZOF and MIL with BP-based object saliency maps using CRF. The detail is explained in Section III-E.

C. Fully convolutional network

"Fully convolutional network (FCN)" originally proposed by Sermanet et al. [9] plays important role in the recent semantic segmentation methods [3], [10]. Fully convolutional network allows an arbitrary-size image by replacing all the full connection layers with 1×1 convolutional layers. In general, the FCN is used with a larger-size image than a usual 256×256 image to obtain a coarse object heatmap as an output. Therefore, with FCN, we can estimate object location directly without a second training step including location estimation such like mi-SVM in Section III-B for weakly-supervised segmentation.

Some works which adopted FCN to weakly-supervised segmentation [15], [17] have been proposed so far. They trained FCNs using neither pixel-wise annotation nor bounding box annotation but image-level annotation with global-max-pooling.

In the proposed FCN-based method, we integrate coarse object heatmaps obtained by FCN and saliency maps the detail of which is explained in Section III-D.

D. BP-based object saliency maps

In [1], they regarded the derivatives of the class scores with respect to an input image as class saliency maps. However, the position of an input image is the furthermost from the class score output on the deep CNN, which sometime causes weakening or vanishing of gradients. Instead of the derivatives of the input image, we use the derivatives of relatively upper intermediate layers which are expected to retain more high-level semantic information. We select the maximum absolute values of the derivatives with respect to the feature maps at each location of feature maps across all the kernels, and up-sample them with bilinear interpolation so that their size becomes the same as an input image. Finally we average them to obtain one saliency map. The idea on aggregating of information extracted from multiple feature layers was inspired by the work of [3], although they extracted not CNN derivatives but feature maps calculated by feed-forwarding.

The class score derivative v_i of the i-th layer is the derivative of class score Sc with respect to the layer L_i at the point (activation signal) L_i:

$$v_i = \left. \frac{\partial S_c}{\partial L} \right|_{L_i} \tag{3}$$

v_i can be computed by back-propagation. After obtained v_i, we up-sample it to w_i with bilinear interpolation so that the size of an 2-D map of v_i becomes the same as an input image. Next, the saliency map $m_{i,x,y}$ is computed as

$$m_{i,x,y} = \max |w_{i,h_i(x,y,k)}| \tag{4}$$

where $h_i(x, y, k)$ is the index of the element of w_i, and k represents kernel. Then, we aggregate $m_{i,x,y}$ for each target layer and obtain a dense saliency maps $g_{x,y}$ are represented as:

$$g_{x,y} = \frac{1}{L} \sum \tanh(\alpha \cdot m_{i,x,y}) \tag{5}$$

where L is the number of layer to aggregate, and α is a scalar constant.

To estimate object saliency maps, we use guided back propagation (GBP) proposed by Springerberg et al. [13] instead of the normal back propagation (BP) used in the work on class saliency map estimation by Simonyan et al. [1]. Only the ways to back propagation through ReLUs (rectified linear units) are different. In the GBP, only positive loss values are propagated back to the previous layers through ReLUs as follows:

$$\text{BP} : \frac{dz^i}{dx^i} = \frac{dz^{i+1}}{dx^{i+1}} \cdot (conv^{i+1} > 0) \tag{6}$$

$$\text{GBP} : \frac{dz^i}{dx^i} = \left(\frac{dz^{i+1}}{dx^{i+1}} > 0 \right) \cdot (conv^{i+1} > 0) \tag{7}$$

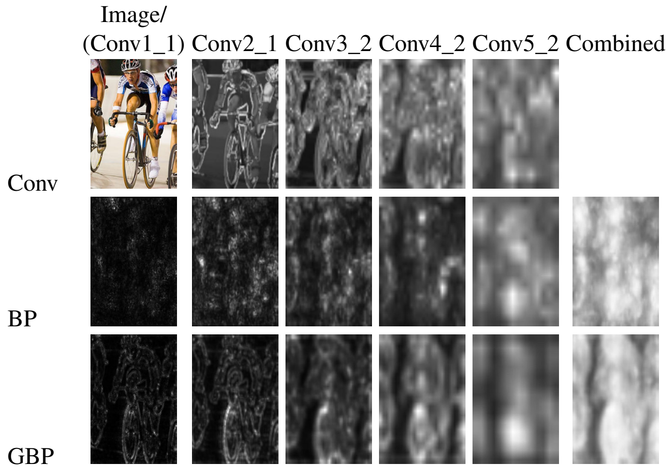

Fig. 3. First row: feature maps (activations) (the given image itself at image-level), Second row: saliency maps by back-propagation (BP), Third row: saliency maps by guided back-propagation (GPB), Columns: image-level, conv2_1, conv3_2, conv4_2, and conv5_2, combined object saliency maps

GBP can emphasize edges of objects, which is a desirable property for estimating object saliency maps. Figure 3 shows up-sampled saliency maps of "bicycle" in the image-level and four intermediate layers of VGG16 [4], w_i, $w_i^{conv2\_1}$, $w_i^{conv3\_1}$, $w_i^{conv4\_1}$, $w_i^{conv5\_1}$, obtained by both BP and guided BP as well as feature maps in case of back-propagating the "bicycle" signal to the network.

By aggregating saliency maps in the intermediate layers, we can obtain more clear object saliency maps. In this paper, we use this object saliency map as generic foreground region prior.

E. Integration of feature maps and saliency maps

In this section, we describe how to integrate feature maps with saliency maps regarding two kinds of the proposed methods, the ZOF-based method and the FCN-based method.

We adopted different integration methods for the two methods. In the ZOF based method, we employed CRF using super-pixels. On the other hand, in the FCN based method, we treated saliency maps as foreground priors which is similar to smoothing priors in [14].

1) CRF with super-pixel for ZOF: Since each zoom-out feature corresponds to each super-pixel, we regard super-pixels as nodes in the CRF graph. We assume that y_p is a label of super-pixel p in Image I, and $\mathbf{y}$ is a aggregated vector of all the y_p. The energy function of CRF is defined as follows:

$$E(\mathbf{y}|I) = \sum_{p \in P} U(y_p|I) + \sum_{p,q \in N} V(y_p, y_q|I) \tag{8}$$

where $U(\cdot)$ is a unary term, and $V(\cdot)$ is a pairwise term. We use as unary potential $U(y_p|I) = -\mathrm{log}Pr(y_p|I)$, where $Pr(y_p|I)$ is the label assignment probability at each super pixel y_p on image I. We obtained the label assignment probability of each object class in foreground by adapting linear SVM which is trained using mi-SVM [18] to zoom-out features. We use saliency maps obtained by back-propagation for estimating background class probability.

We define a pairwise term as follows referring to [19], [20], [21]:

$$V(y_p, y_q|I) = \left(\frac{L(p,q)}{1 + \|p - q\|}\right) [y_p \neq y_q] \tag{9}$$

where $\|p - q\|$ is a distance between super-pixel p and q regarding LUV color vectors, and $L(p,q)$ is the length of the boundaries shared by super-pixel p and q.

2) Saliency maps as smoothing prior for FCN: We enhance coarse object heatmaps which are obtained as the final layer output of FCN by multiplying saliency maps. We up-sample coarse object heatmaps and saliency maps to unify their size in advance. Where $f_{x,y}$ represents coarse object heatmaps and $g_{x,y}$ represents saliency maps at pixel (x, y), the segmentation result $h_{x,y}$ is obtained as follows:

$$h_{x,y} = \begin{cases} k, & \text{if } \arg\max_{c \in C} f^c_{x,y} g_{x,y} > \delta \\ k_{bg} & \text{otherwise} \end{cases} \tag{10}$$

C is a set of the target object classes, and δ is a pre-defined threshold.

IV. Experiments

A. Dataset

In the experiments, we use the PASCAL VOC 2012 segmentation data [22] to evaluate the proposed methods. The PASCAL VOC dataset consists of 1464 training images, 1449 validation images, and 1456 test images including 20 class pixel-labels. In addition, we used additional PASCAL 20 class data including 10582 train_aug images provided by Hariharan et al. [23].

B. Experimental setup

1) Training of CNN: We used 16-layered CNN, VGG-16 [4] pre-trained with ImageNet 1000 categories as a basic CNN architecture. We fine-tuned VGG-16 using PASCAL VOC training dataset and train_aug by Hariharan et al [23] with Sigmoid entropy loss for multi-label training as described in section III-A with batch size 16 and learning rate 1e-5, momentum 0.9 and weight decay 0.0005. For the first 30000 iterations, we fine-tuned only the upper layers of the modified VGG-16 than Pool_5, and for the next 20000 iterations, we fine-tuned all the layers.

2) Zoom-out features: In the ZOF-based method, we extracted about 500 super-pixels by the SLIC super-pixels [24] from all the training images, and calculated Zoom-out features (ZOF) [2]. While in [2] they extracted ZOF from all the layers of VGG-16, 13 convolutional layers and 3 fully connected layers, we extracted ZOF from 13 convolutional layers, pool5 and fc7. When applying mi-SVM [18] for each class, we used 500 images of the target class as positive samples and 1000 images of the other classes than the target class. We used the classification result of the CNN to limit the possible object classes for CRF.

3) Fully convolutional networks: In the FCN based method, we used image-level recognition results as image-level prior (ILP) for post processing which is noted by [14] to consider global context. Specifically, we applied global-max-pooling to the obtain heatmaps to obtain ILP in the same way as the training phase, and multiplied each class pooled score and each class heatmap value.

Our approach differs from [14] on the method of enhancing coarse heatmaps. Pedro et al. [14] used Multi-scale Combinatorial Grouping (MCG) [25] which was known as a region proposal method for correcting heatmaps which is called as "smoothing priors (SP)" and made foreground mask by aggregating objectness scores of about 2000 region candidates. On the other hand, we used saliency maps obtained by back-propagation and corrected coarse heatmaps in similar way to MCG smoothing priors. Then, we compared our method with MCG smoothing priors in the experiments.

4) Saliency maps: Back-propagation needs higher computational cost than feed-forward computation. Thus, we computed backward once for a given image even if there are several class objects. We predicted presence/absence of objects in the image by feed-forwarding, and made a back-propagating signal which is the same dimension to the output vector of the CNN. Simply, we set the signal on values is 1 for presence classes and is 0 for absence classes. We propagated the signal by back-propagation from the top convolutional layer, extracted gradients from the layer conv3_2 and conv4_2 and conv5_2, and aggregated them following the method explained in Section III-D.

C. Experimental Results

Table I and II shows the results on the proposed methods and some other state-of-the-art methods for PASCAL VOC 2011 validation data and PASCAL VOC 2012 test data. Note that although [14] showed the high performance, they used 700,000 additional training images selected from ImageNet which about 70 times as large as the common additional training data [23]. We report three results by our proposed methods, and compare them with other state-of-the-art weakly supervised segmentation methods. "ZOF with GBP" and "FCN with GBP" means the proposed methods which integrate feature maps and saliency maps obtained by guided back-propagation (GBP). We also report the "FCN with MCG" result to compare effect of saliency maps with smoothing priors of "MIL-seg" [14] which is generated by MCG. Therefore, we examine the effectiveness of GBP-based object saliency maps as foreground priors against the smoothing priors [14] by comparing "FCN with GBP" and "FCN with MCG".

"ZOF with GBP" achieved the better or near to result of state-of-the-art methods. "FCN with GBP" outperformed MIL-FCN [15], EM-Adapt [17], CCNN [16] using only train_aug samples which were provided by Hariharan et al. [23] on validation set and test set. "FCN with GBP" also outperformed MIL-ILP-seg [14] which used additional images on test set. We shows some example results in Figure 4.

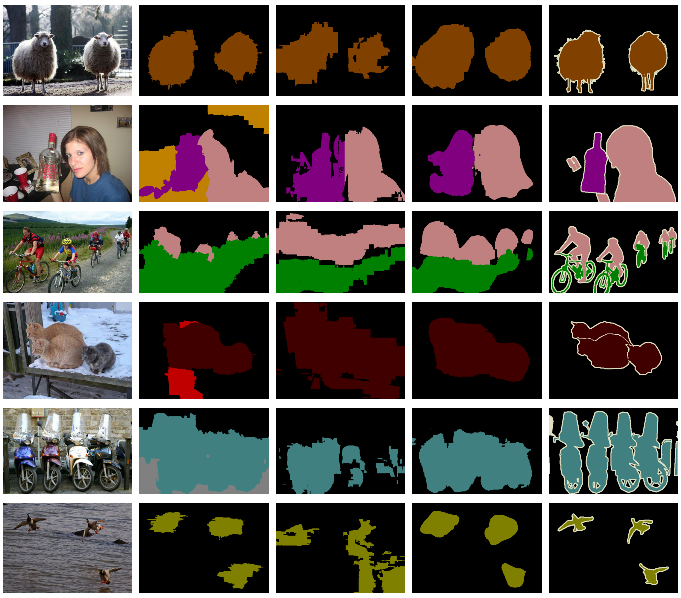

Fig. 4. For each row, we show the input image, result of ZOF with GBP, and FCN with MCG, and FCN with GBP, and ground truth label.

We also compared "FCN with GBP" with "FCN with MCG" to verify the effect of saliency map based and MCG-based smoothing priors. As a result, "FCN with GBP" outperformed "FCN with MCG" clearly, i.e.,33.8% vs. 41.4% (val. set), 33.1% vs. 40.7% (test set). Therefore, combining saliency maps obtained by guided back-propagation and feature maps of CNN is the most effective for the weakly supervised segmentation task. Figure 5 shows comparison between saliency maps and MCG priors, which exhibits that saliency maps reacted to only target object regions more correctly.

V. Conclusions

In this paper, we proposed a novel weakly-supervised segmentation method which were based on feature maps and back-propagation-based object saliency maps [1]. In the proposed method, we showed that denser and clearer saliency map can be obtained by up-sampling saliency maps of the intermediate layers and aggregating them. We achieved the state-of-the-arts in the weakly supervised segmentation task, and confirmed the effectiveness of back-propagation-based object saliency maps as smoothing priors.

References

[1] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," in *Proc. of International Conference on Learning Representations*, 2014.

[2] M. Mostajabi, P. Yadollahpour, and G. Shakhnarovich, "Feedforward semantic segmentation with zoom-out features," in *Proc. of IEEE Computer Vision and Pattern Recognition*, 2015.

[3] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. of IEEE Computer Vision and Pattern Recognition*, 2015.

[4] K. Simonyan, A. Vedaldi, and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. of International Conference on Learning Representations*, 2015.

[5] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. of IEEE Computer Vision and Pattern Recognition*, 2014.

TABLE I
RESULTS ON PASCAL VOC 2011 VALIDATION SET. (NOTE THAT MIL-SPPXL/BB/SEG* USED ADDITIONAL LARGE TRAINING SETS.)

Methods	bg	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	mean
MIL-FCN [15]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	25.7
EM-Adapt [17]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	38.2
CCNN [16]	65.9	23.8	17.6	22.8	19.4	36.2	47.3	46.9	47.0	16.3	36.1	22.2	43.2	33.7	44.9	39.8	29.9	33.4	22.2	**38.8**	36.3	34.5
MIL-sppxl* [14]	77.2	37.3	18.4	25.4	28.2	31.9	41.6	48.1	50.7	12.7	45.7	14.6	50.9	44.1	39.2	37.9	28.3	44.0	19.6	37.6	35.0	36.6
MIL-bb* [14]	78.6	46.9	18.6	27.9	30.7	38.4	44.0	49.6	49.8	11.6	44.7	14.6	50.4	44.7	40.8	38.5	26.0	45.0	20.5	36.9	34.8	37.8
MIL-seg* [14]	**79.6**	**50.2**	21.6	**40.6**	**34.9**	**40.5**	45.9	**51.5**	**60.6**	12.6	**51.2**	11.6	**56.8**	**52.9**	44.8	42.7	31.2	**55.4**	21.5	**38.8**	**36.9**	**42.0**
ZOF with GBP (ours)	70.6	44.4	24.7	37.5	16.4	33.3	**60.6**	35.5	58.8	5.5	45.5	15.9	53.4	41.1	**54.8**	39.6	24.2	52.1	18.4	38.6	27.5	38.1
FCN with MCG (ours)	71.0	21.9	18.5	22.0	12.8	34.6	37.5	43.3	47.1	17.7	38.5	29.4	40.9	43.3	40.7	38.7	29.0	35.6	22.8	36.6	27.5	33.8
FCN with GBP (ours)	76.8	40.0	**28.1**	38.6	24.6	39.7	37.3	50.2	51.4	**23.9**	47.2	**25.8**	53.6	49.1	53.9	**45.1**	**36.0**	48.1	**30.0**	35.8	33.9	41.4

TABLE II
RESULTS ON PASCAL VOC 2012 TEST SET. (NOTE THAT MIL-IPL-SPPXL/BB/SEG* USED ADDITIONAL LARGE TRAINING SETS.)

Methods	bg	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	mean
MIL-FCN [15]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	24.9
EM-Adapt [17]	76.3	37.1	21.9	41.6	26.1	38.5	50.8	44.9	48.9	**16.7**	40.8	29.4	47.1	45.8	54.8	28.2	30.0	44.0	**29.2**	**34.3**	46.0	39.6
CCNN [16]	-	21.3	17.7	22.8	17.9	38.3	51.3	43.9	51.4	15.6	38.4	17.4	46.5	38.6	53.3	40.6	**34.3**	36.8	20.1	32.9	38.0	35.5
MIL-ILP-sppxl* [14]	74.7	38.8	19.8	27.5	21.7	32.8	40.0	50.1	47.1	7.2	44.8	15.8	49.4	47.3	36.6	36.4	24.3	44.5	21.0	31.5	41.3	35.8
MIL-ILP-bb* [14]	76.2	42.8	20.9	29.6	25.9	38.5	40.6	51.7	49.0	9.1	43.5	16.2	50.1	46.0	35.8	38.0	22.1	44.5	22.4	30.8	43.0	37.0
MIL-ILP-seg* [14]	**78.7**	48.0	21.2	31.1	**28.4**	35.1	51.4	**55.5**	52.8	7.8	**56.2**	19.9	**53.8**	50.3	40.0	38.6	27.8	51.8	24.7	33.3	**46.3**	40.6
ZOF with GBP (ours)	71.1	**48.4**	24.4	**48.5**	15.2	38.2	**65.6**	32.8	**57.9**	5.1	43.8	18.2	46.2	48.7	50.4	35.7	22.5	41.7	19.1	29.2	27.1	37.7
FCN with MCG (ours)	71.9	21.8	18.4	25.4	14.9	35.2	40.0	39.7	41.5	13.4	36.4	**29.9**	36.5	45.4	41.3	38.7	26.9	34.5	19.7	29.8	33.3	33.1
FCN with GBP (ours)	78.0	35.8	**28.5**	45.7	25.9	**43.1**	40.1	46.9	49.1	16.3	42.4	29.6	50.8	**51.3**	**57.2**	**44.4**	28.9	**44.8**	27.5	31.6	36.2	**40.7**

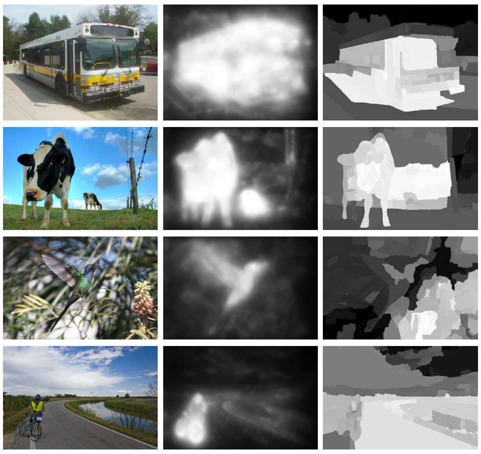

Fig. 5. For each row, (left) image, (middle)saliency maps, (right)MCG priors.

[6] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik, "Simultaneous detection and segmentation," in *Proc. of European Conference on Computer Vision*, 2014.

[7] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," *International Journal of Computer Vision*, vol. 104, no. 2, pp. 154–171, 2013.

[8] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Is object localization for free? -weakly-supervised learning with convolutional neural networks," in *Proc. of IEEE Computer Vision and Pattern Recognition*, 2015.

[9] P. Sermanet, D. Eigen, X. Zhang, M.l Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," in *Proc. of International Conference on Learning Representations*, 2014.

[10] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and Yuille A. L., "Semantic image segmentation with deep convolutional nets and fully connected CRFs," in *Proc. of International Conference on Learning Representations*, 2015.

[11] S. Zheng, S. Jayasumana, B. R. Paredes, V. Vineet, and Z. Su, "Conditional random fields as recurrent neural networks," in *Proc. of IEEE Computer Vision and Pattern Recognition*, 2015.

[12] S. Bell, P. Upchurch, N. Snavely, and K. Bala, "Material recognition in the wild with the materials in context database," in *Proc. of IEEE Computer Vision and Pattern Recognition*, 2015.

[13] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," in *Proc. of International Conference on Learning Representations*, 2015.

[14] P. Pedro and C. Ronan, "From image-level to pixel-level labeling with convolutional networks," in *Proc. of IEEE Computer Vision and Pattern Recognition*, 2015.

[15] D. Pathak, E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional multi-class multiple instance learning," in *Proc. of International Conference on Learning Representations*, 2015.

[16] D. Pathak, P. Krahenbuhl, and T. Darrell, "Constrained convolutional neural networks for weakly supervised segmentation," in *Proc. of IEEE International Conference on Computer Vision*, 2015.

[17] G. Papandreou, L.-C. Chen, K. Murphy, and A. L. Yuille, "Weakly-and semi-supervised learning of a dcnn for semantic image segmentation," in *Proc. of IEEE International Conference on Computer Vision*, 2015.

[18] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," in *Advances in Neural Information Processing Systems*, 2002.

[19] Y. Y. Boykov and M.-P. Jolly, "Interactive graph cuts for optimality boundary & region segmentation of objects in n-d images," in *Proc. of IEEE International Conference on Computer Vision*, 2001.

[20] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation," in *Proc. of European Conference on Computer Vision*, 2006.

[21] B. Fulkerson, A. Vedaldi, and S. Soatto, "Class segmentation and object localization with superpixel neighborhoods," in *Proc. of IEEE International Conference on Computer Vision*, 2009.

[22] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. 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.

[23] B. Hariharan, P. Arbelaez, L. Bourdev, S. Maji, and Malik. J., "Semantic contours from inverse detectors," in *Proc. of IEEE International Conference on Computer Vision*, 2011.

[24] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Ssstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2012.

[25] A. Pablo, Jonathan T. Jordi, P., M. Ferran, and M. Jitendra, "Multiscale combinatorial grouping," in *Proc. of IEEE Computer Vision and Pattern Recognition*, 2014.