# IMPROVING THE DISCRIMINATION BETWEEN FOREGROUND AND BACKGROUND FOR SEMANTIC SEGMENTATION

*Yu Liu and Michael S. Lew*

Department of Computer Science, Leiden University, Netherlands

## ABSTRACT

One challenging problem in semantic segmentation is due to the erroneous predictions between categorical foreground and cluttered background. To address it, we propose to utilize a fused loss function to train a fully convolutional network, which aims to enhance the discrimination between foreground and background in images. In addition, we propose a pixel objectness (POS) to measure the importance of pixels. POS is able to recover some missing foreground pixels from the background. Experimental results on the PASCAL VOC 2012 dataset demonstrate our approach can achieve considerable improvements compared with the baseline counterpart, while maintaining the ease of training deep networks.

***Index Terms***— Semantic Segmentation, Fully Convolutional Networks, Fused Loss Function, Pixel Objectness

## 1. INTRODUCTION

Semantic segmentation, which aims to assign a pre-defined object label for each image pixel , has been a fundamental task in computer vision research [1–3]. Inspired by the significant success from convolutional neural networks (CNN) [4–6], a number of works [7–13] have applied CNNs to semantic segmentation, and yielded state-of-the-art performance.

In the following, we summarize recent semantic segmentation approaches based on CNNs in three aspects. (1) *Detection-based segmentation* [6–9, 14]: segment images based on the candidate windows predicted by object detection approaches. (2) *FCN-based segmentation* [10,11,15–18]: build fully convolutional networks (FCN) by replacing the fully-connected layers with more convolutional layers. FCN is well-suited for pixel-level classification. (3) *Weakly supervised segmentation* [19–25]: use weak annotations (e.g. image-level or window-level labels) to train the segmentation networks. Although its performance has increased rapidly in recent years, semantic segmentation still remains a challenging task, especially for small foreground objects and cluttered background. One issue is due to erroneously classifying pixels between foreground classes (*i.e.*pre-defined objects) and background class. In particular, some object pixels that are not quite salient might be predicted as background classes. Also, the background pixels similar to objects might be assigned with object labels. Motivated by these observation, in this work, we primarily focus on the incorrect predictions between foreground classes and background class, which have not been studied much in prior work. To be more specific, we use a fused loss function while training a FCN, to explicitly discriminate the differences between various foreground classes and one background class. Additionally we propose to add a new measurement, called pixel objectness, to improve the unary potential in conditional random fields (CRFs).

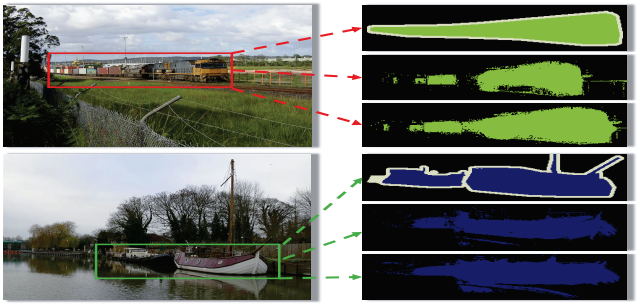


**Fig. 1**: Example of two segmentation results. The left side shows the input images and a window is selected to highlight the salient object. The right side (from top to bottom) lists the segmentation masks from the ground-truth, the baseline, and our approach. For both the train and boat, our approach recovers more foreground pixels from background than the baseline.

Our main contributions are summarized as follows:

(1) Apart from using a standard softmax loss function, we also adapt a positive-sharing loss function [26]. We integrate the two loss functions together and fuse them to train the FCN model jointly. The fused loss can not only preserve the discrimination among foreground classes, but also penalize the mistakes between foreground and background.

(2) In addition, we propose a method to measure the pixel importance (called pixel objectness) in one image, called pixel objectness (POS). and add it to the unary potential in CRFs. Especially, we assign POS to probability maps of object classes, but not background class. This approach can help recover useful foreground pixels from the background.

(3) Compared with the FCN baseline [10], our approach obtains considerable improvements in both IoU and recall performance measurements, especially regarding small details about objects. Figure 1 compares the segmentation results between our approach and the baseline.

## 2. METHODOLOGY

This section will detail our approach with respect to alleviate the incorrect classification between foreground and background. Figure 2 illustrates the overview of the approach.

### 2.1. Problem Formulation

Assume that there is a training dataset which contains $N$ images: $\{x^{(i)}, y^{(i)}\}_{i=1}^{N}$, where $x^{(i)}$ is the $i$-th input image and $y^{(i)}$ is its ground-truth segmentation mask. $x^{(i)}$ has $M^{(i)}$ pixels, where $M^{(i)} = H^{(i)} \cdot W^{(i)}$, and $H^{(i)}$ and $W^{(i)}$ is the image height and width. Let $x_j^{(i)}$ be the $j$-th raw pixel, $j = 1, \ldots, M^{(i)}$. Assume that there are one background class and $C$ pre-defined object classes. $y_j^{(i)} \in \{0, 1, \ldots, C\}$ is the ground-truth label of $x_j^{(i)}$. If $y_j^{(i)} > 0$, then $x_j^{(i)}$ is a object pixel; Otherwise, $x_j^{(i)}$ is a background pixel.

### 2.2. Fused Loss Function for FCN

Given an input image, FCN [10] allows to infer a number of probability maps in the topmost layer. Every map measures the probability about one pixel belonging to the corresponding object class. We denote $s_{j,k}^{(i)}$ as feature activations in the last layer of FCN model, where $(k = 0, 1, \ldots, C)$. Then the probability that $x_j^{(i)}$ belongs to the $k$-th class is computed by the softmax normalization

$$p_{j,k}^{(i)} = \frac{\exp(s_{j,k}^{(i)})}{\sum_{l=0}^{C} \exp(s_{j,l}^{(i)})}. \tag{1}$$

Then, the FCN model calculates the softmax loss cost $\mathcal{L}_0$ by the following formulation

$$\mathcal{L}_0 = -\frac{1}{N}\left[\sum_{i=1}^{n}\sum_{j=1}^{M_i}\sum_{k=0}^{C} \mathbf{h}(y_j^{(i)} = k) \log p_{j,k}^{(i)}\right], \tag{2}$$

where $\mathbf{h}(i = j)$ is the Kronecker delta response $\delta_{i,j}$.

We can find that the standard softmax loss equally computes the classification error for each class. However, much error in semantic segmentation is attributed to the incorrect predictions between foreground and background. To be more specific, some foreground pixels belonging to object classes are incorrectly predicted as background, and vice verse. Hence, it is significantly essential to penalize these error. In [26], they converted two-class contour detection to a multi-class classification task, and presented a positive-sharing loss function that was fused with the standard softmax loss. Inspired by them but different, we aim to convert the multi-class segmentation problem to a two-class classification task. We treat the background as a negative class, and integrate all objects into a positive class. Therefore, we can

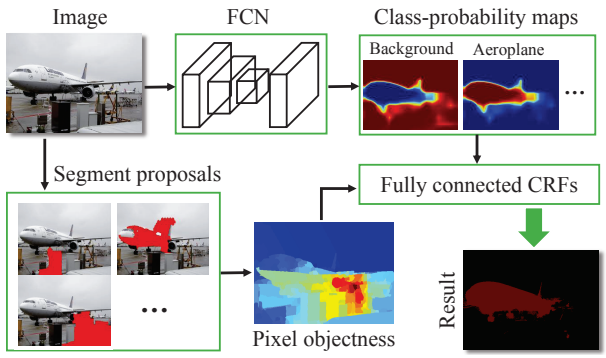



**Fig. 2**: Overview of our semantic segmentation approach. First, we use the fused loss function to train a FCN model, which can predict a number of class-probability maps. Then, we compute pixel objectness based on segment proposals for all pixels in the input image. Finally, the probability maps and POS are integrated into the CRFs to jointly estimate the segmentation mask.

compute the positive-sharing loss between the positive and negative class by

$$\mathcal{L}_1 = -\frac{1}{N}\left[\sum_{i=1}^{N}\sum_{j=1}^{M_i}\Big(\mathbf{h}(y_j^{(i)} = 0)\log p_{j,0}^{(i)} + \sum_{k=1}^{C}\mathbf{h}(y_j^{(i)} = k)\log(1 - p_{j,0}^{(i)})\Big)\right]. \tag{3}$$

As all objects in the positive class compute the same loss cost, the positive-sharing loss function is able to increase the discrimination between foreground and background. Moreover, we fuse the standard softmax function and the positive-sharing loss function together. The fused loss function can not only retain the separability of each class, but also penalize the error between foreground and background. The formulation is expressed by $\mathcal{L} = W_s \cdot \mathcal{L}_0 + W_p \cdot \mathcal{L}_1$, where $\mathcal{L}$ is the fused loss; $W_s$ and $W_p$ are the weights to balance the standard softmax loss and positive-sharing loss.

During the training procedure, we compute the partial derivatives of the fused loss *w.r.t* parameters, and update them by SGD [4]. The partial derivatives of the loss $\mathcal{L}$ *w.r.t* $s_{j,0}^{(i)}$ and $s_{j,l}^{(i)}, l = 1, \ldots, C$, are formulated via

$$\frac{\partial \mathcal{L}}{\partial s_{j,0}^{(i)}} = \frac{1}{N}\Big[(W_s + W_p)\mathbf{h}(y_j^{(i)} = 0)(p_{j,0}^{(i)} - 1) + (W_s + W_p)\sum_{k=1}^{C}\mathbf{h}(y_j^{(i)} = k)p_{j,0}^{(i)}\Big], \tag{4}$$

$$\frac{\partial \mathcal{L}}{\partial s_{j,l}^{(i)}} = \frac{1}{N}\left[\Big(W_s + W_p\mathbf{h}(y_j^{(i)} = 0)\Big)p_{j,l}^{(i)} - W_s\mathbf{h}(y_j^{(i)} = l) - W_p\sum_{k=1}^{C}\mathbf{h}(y_j^{(i)} = k)\Big(\frac{p_{j,0}^{(i)}p_{j,l}^{(i)}}{1 - p_{j,0}^{(i)}}\Big)\right]. \tag{5}$$

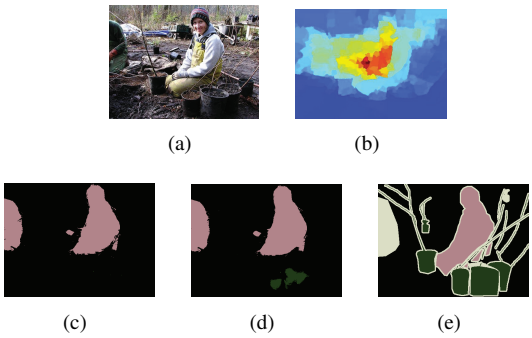


**Fig. 3**: (a) Input image; (b) Pixel objectness map; (c) CRFs without POS; (d) CRFs with POS; (e) Ground-truth.

### 2.3. Pixel Objectness for CRFs

CRFs are often used as a post-processing stage for detailed boundary recovery [11, 15]. Each pixel represented with one probability vector from FCNs can be fed into CRFs to compute the unary potential. However, some important object pixels may be labeled as background due to their weak unary potentials. To increase the unary potentials of object classes, we propose a pixel objectness (POS) measurement, which can recover more useful object pixels from the background.

#### *2.3.1. Pixel objectness*

Extracting segment proposals [27, 28] has become an efficient preprocessing stage for the folowing segmentation [7, 16]. The objectness of a proposal or window measures its probability of being one salient object. Instead, we compute a new measurement with respect to pixels, called pixel objectness (POS). POS indicates the probability about one pixel locating within one salient object. Our hypothesis is that if there are more proposals containing one pixel, then this pixel should be assigned with a larger weight (or objectness).

First, we use the geodesic object proposals (GOP) [27] to extract segment proposals. It identifies critical level sets in geodesic distance transforms that are computed for seeds in the image. GOP can achieve promising accuracy at a fraction of the computational cost. We choose the learned GOP(140,4) method that can generate about 700 segment proposals in about 1 second. Assume that an image $x^{(i)}$ has $G^{(i)}$ segment proposals in total. For each pixel, we count how many proposals contain this pixel, denoted as $g_j^{(i)}, j = 1, \ldots, M^{(i)}$. Finally, POS is calculated by

$$pos_j^{(i)} = \frac{g_j^{(i)}}{G^{(i)}}, pos_j^{(i)} \in [0, 1]. \tag{6}$$

In Fig. 3(b), we illustrate the $pos_j^{(i)}$ map for one image. The person is the most salient part in the image, and the plant below also contains larger objectness than the background.

#### *2.3.2. Unary potential with pixel objectness*

The energy function of CRFs [29] is represented by

$$E(x) = \sum_{j}^{M^{(i)}} \theta(x_j^{(i)}) + \sum_{j_1}^{M^{(i)}} \sum_{j_2}^{M^{(i)}} \theta(x_{j_1}^{(i)}, x_{j_2}^{(i)}), \tag{7}$$

where $\theta(x_j^{(i)})$ is the unary potential at pixel $x_j^{(i)}$. $\theta(x_{j_1}^{(i)}, x_{j_2}^{(i)})$ represents the pairwise potential. In [11, 19], the unary potential was computed with the prediction probabilities from FCN. Different from them, we add POS to the computation of the unary potential as follows:

$$\theta_k(x_j^{(i)}) = \begin{cases} -\log p_{j,0}^{(i)} & k = 0 \\ -\log \left( p_{j,k}^{(i)} \cdot \exp(pos_j^{(i)}) \right), & 1 \leq k \leq C \end{cases} \tag{8}$$

where $\theta_k(x_j^{(i)}) \in \theta(x_j^{(i)}), k = (0, \ldots, C)$; $p_{j,k}^{(i)}$ is the prediction as seen in Equation (1). This new unary potential elevates the probability of object classes with a rate of $\exp(pos_j^{(i)})$. When some of object pixels already have low POS scores, the rate will be close to 1 and therefore their prediction probabilities remain nearly unchanged. However, when they have large POS scores, the increased rate is able to compensate for their prediction probabilities. It is worthy mentioning that we do not add POS to the unary potential of background pixels. In a word, POS allows to avoid some important object pixels to be classified as background. Similarly, we compute the pairwise potential $\theta(x_{j_1}^{(i)}, x_{j_2}^{(i)})$ with bilateral position and color intensities [11, 29].

We present one comparison example in Fig. 3(c)(d). CRFs without POS can not segment any plant. However, CRFs with GOP is capable of recovering some plant below the person.

## 3. EXPERIMENTS

We evaluate the performance of our approach and compare it with the FCN baseline [10].

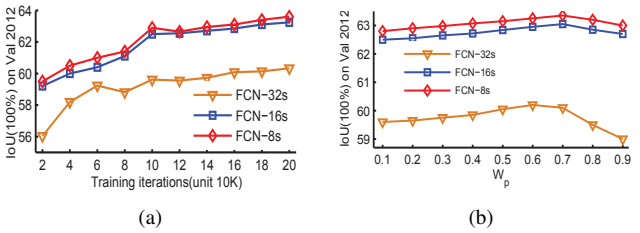
### 3.1. Dataset

The PASCAL VOC 2012 segmentation dataset [30] consists of 20 foreground object classes and one background class. The original dataset contains 1464 training images, 1449 validation images, and 1456 test images. When evaluating the val set, we use a merged training dataset including the original training images and the augmented training images [31]. As there are validation images included in the merged training set, we pick the non-intersecting set of 904 images and. The performance is measured in terms of pixel intersection-over-union (IoU) measurement, and recall measurement.

We add CRFs inference [11] to FCN predictions [10], including FCN-32s, FCN-16s and FCN-8s. These baselines are named with *SoftmaxLoss+CRFs*. In comparison, Our method is called as *FusedLoss+POS-CRFs*.

**Table 1**: 20 object classes results on the PASCAL VOC 2012 val set (better results in bold).

| methods | mean | aero | bike | bird | boat | bottle | bus | car | cat | chair | cow | table | dog | horse | mbike | person | plant | sheep | sofa | train | tv |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| FCN [10] | 62.2 | 76.8 | **34.2** | 68.9 | 49.4 | 60.3 | 75.3 | 74.7 | 77.6 | 21.4 | 62.5 | 46.8 | 71.8 | 63.9 | **76.5** | 73.9 | 44.9 | 72.4 | 37.4 | 70.9 | 55.1 |
| Baseline | 65.85 | 79.9 | 32.0 | **83.1** | 54.2 | 67.5 | 80.0 | 72.5 | 81.1 | **27.2** | 72.5 | 53.2 | 74.1 | **69.1** | 74.4 | 77.2 | 44.3 | 73.9 | 42.6 | 75.8 | **57.6** |
| Ours | **66.71** | **81.5** | 33.7 | 81.7 | **56.8** | **68.4** | **81.1** | **76.4** | **81.3** | 26.8 | **73.2** | **54.1** | **75.0** | **69.1** | 75.5 | **78.0** | **45.1** | **75.0** | **43.3** | **76.1** | 57.3 |




(a) (b)

**Fig. 4**: (a) PASCAL VOC val set accuracy with training iterations ranging from 20K to 200K. (b) Weights analysis: $W_p$ is the weight of the positive-sharing loss.

### 3.2. Implementation details

Our implementation is based on the Caffe library [32]. Following [10], we used a fixed learning rate of $10^{-4}$, a weight decay of 0.0001, a momentum of 0.9, and a mini-batch size of 1. We reproduced the FCN baseline results on VOC 2012 val set as seen in Fig. 4(a). After 100K training iterations, our results based on FCN-32s (59.61%), FCN-16s (62.52%), FCN-8s (62.91%) are competitive with [10]. Therefore, we train models with 100K iterations in the following experiments.

### 3.3. Analysis on weights

In this experiment, we evaluate the importance of two loss functions in the fused loss ($W_p + W_s = 1$). We change $W_p$ from 0.1 to 0.9, and compute the IoU on the val set. The results are presented in Fig. 4(b). As for FCN-32s, the superior accuracy is obtained when $W_p = 0.6$. Likewise, when $W_p = 0.7$, FCN-16s and FCN-8s can get the best performance. This evaluation validates that the fused loss is beneficial for improving the performance. In the following, we employ these optimal weights for other experiments.

### 3.4. Comparison with the baseline

We evaluate the methods with two measurements as follows:

**IoU measurement**. As reported in Table 2, we implement FCN-32s, FCN-16s, FCN-8s with different configurations. First, the fused loss increases about 0.45-0.6% accuracy, compared with the standard softmax loss, as seen in the first and second row. Specifically, FCN-8s with the fused loss yields 63.35% IoU. Second, the standard CRFs in [11] can boost the accuracy with remarkable improvements, see the results in the third row and the forth row. Third, when we apply the POS to the CRFs, the models can get about 0.3% IoU gain as compared with the standard CRFs. The results are shown in the fifth and sixth row. In summary, our segmentation method (FusedLoss+POS-CRFs) can achieve about 1% boost, compared with the counterpart baseline (SoftmaxLoss+CRFs). In addition, we give detailed comparison on the 20 objects classes in Table 1. For most classes, our method (FCN-8s+FusedLoss+POS-CRFs) is better than the baseline (FCN-8s+SoftmaxLoss+CRFs).

**Table 2**: IoU measurement on the VOC 2012 val set.

| Method | FCN-32s | FCN-16s | FCN-8s |
|---|---|---|---|
| SoftmaxLoss | 59.61 | 62.52 | 62.91 |
| FusedLoss | 60.22 | 63.05 | 63.35 |
| SoftmaxLoss+CRFs | 62.64 | 65.45 | 65.85 |
| FusedLoss+CRFs | 63.21 | 66.05 | 66.42 |
| SoftmaxLoss+POS-CRFs | 62.95 | 65.8 | 66.15 |
| FusedLoss+POS-CRFs | **63.55** | **66.42** | **66.71** |

**Table 3**: Recall measurement on the VOC 2012 val set.

| Method | FCN-32s | FCN-16s | FCN-8s |
|---|---|---|---|
| Baseline | 68.65 | 72.58 | 74.98 |
| Ours | **70.84** | **74.71** | **77.15** |

**Recall measurement**. Our main motivation is to recover some useful object pixels from the background. To measure the effectiveness of our method, it is essential to compute the recall rates of object pixels. Note that we merge 20 objects as one foreground class. The recall of the foreground class is computed by $\frac{\#correct}{\#total}$, where $\#total$ is the number of object pixels in one image and $\#correct$ indicates how many object pixels are detected correctly. In Table 3, our methods yield modest recall improvements as compared with the baseline.

## 4. CONCLUSIONS

To reduce incorrect predictions between objects and background in semantic segmentation, we exploited a fused loss function for training FCN. Furthermore, we proposed to compute the POS measurement to increase the weight of object pixels, and applied POS to the unary potential in CRFs. Experimental results demonstrates that the proposed approach obtains modest overall accuracy gains, and it can recover more boundary details from the background. Our approach is flexible to be applied to other state-of-the-art methods.

**Acknowledgments** This work was supported by the LIACS Media Lab at Leiden University and the China Scholarship Council. We would like to thank NVIDIA for the donation of GPU cards.

## 5. REFERENCES


[1] Pablo Arbelaez, Michael Maire, Charless Fowlkes, and Jitendra Malik, "Contour detection and hierarchical image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 898–916, 2011.

[2] Pablo Arbelaez, Bharath Hariharan, Chunhui Gu, Saurabh Gupta, and Jitendra Malik, "Semantic segmentation using regions and parts," in *CVPR*, 2012.

[3] Eva Mohedano, Graham Healy, Kevin McGuinness, Xavier Giró-i Nieto, Noel E. O'Connor, and Alan F. Smeaton, "Object segmentation in images using eeg signals," in *22Nd ACM International Conference on Multimedia*, 2014.

[4] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012.

[5] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.

[6] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *CVPR*, 2014.

[7] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik, "Simultaneous detection and segmentation," in *ECCV*, 2014.

[8] Bharath Hariharan, Pablo Andrés Arbeláez, Ross B. Girshick, and Jitendra Malik, "Hypercolumns for object segmentation and fine-grained localization," in *CVPR*, 2015.

[9] Jifeng Dai, Kaiming He, and Jian Sun, "Convolutional feature masking for joint object and stuff segmentation," in *CVPR*, 2015.

[10] Jonathan Long, Evan Shelhamer, and Trevor Darrell, "Fully convolutional networks for semantic segmentation," in *CVPR*, 2015.

[11] 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.

[12] Guo-Jun Qi, "Hierarchically gated deep networks for semantic segmentation," in *CVPR*, 2016.

[13] Mahdyar Ravanbakhsh, Hossein Mousavi, Moin Nabi, Mohammad Rastegari, and Carlo Regazzoni, "Cnn-aware binary map for general semantic segmentation," in *ICIP*, 2016.

[14] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari, "Region-based semantic segmentation with end-to-end training," in *ECCV*, 2016.

[15] Guosheng Lin, Chunhua Shen, Ian Reid, and Anton van dan Hengel, "Efficient piecewise training of deep structured models for semantic segmentation," in *CVPR*, 2016.

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

[17] Ziwei Liu, Xiaoxiao Li, Ping Luo, Chen-Change Loy, and Xiaoou Tang, "Semantic image segmentation via deep parsing network," in *ICCV*, 2015.

[18] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han, "Learning deconvolution network for semantic segmentation," in *ICCV*, 2015.

[19] George Papandreou, Liang-Chieh Chen, Kevin Murphy, and Alan L. Yuille, "Weakly- and semi-supervised learning of a deep convolutional network for semantic image segmentation," in *ICCV*, 2015.

[20] Jifeng Dai, Kaiming He, and Jian Sun, "Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation," in *ICCV*, 2015.

[21] Deepak Pathak, Philipp Krahenbuhl, and Trevor Darrell, "Constrained convolutional neural networks for weakly supervised segmentation," in *ICCV*, 2015.

[22] Yuhang Wang, Jing Liu, Yong Li, and Hanqing Lu, "Semi- and weakly- supervised semantic segmentation with deep convolutional neural networks," in *ACM International Conference on Multimedia*, 2015.

[23] Xiwen Yao, Junwei Han, Gong Cheng, and Lei Guo, "Semantic segmentation based on stacked discriminative autoencoders and context-constrained weakly supervised learning," in *ACM International Conference on Multimedia*, 2015.

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

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

[26] Wei Shen, Xinggang Wang, Yan Wang, Xiang Bai, and Zhijiang Zhang, "DeepContour: A deep convolutional feature learned by positive-sharing loss for contour detection," in *CVPR*, 2015.

[27] Philipp Krähenbühl and Vladlen Koltun, "Geodesic object proposals," in *ECCV*, 2014.

[28] Pablo Arbelaez, J. Pont-Tuset, Jon Barron, F. Marqués, and Jitendra Malik, "Multiscale combinatorial grouping," in *CVPR*, 2014.

[29] Philipp Krähenbühl and Vladlen Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," in *NIPS*, 2011.

[30] 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.

[31] Bharath Hariharan, Pablo Arbelaez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik, "Semantic contours from inverse detectors," in *ICCV*, 2011.

[32] 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 Multimedia*, 2014.