Combining Bottom-Up, Top-Down, and Smoothness Cues for Weakly Supervised Image Segmentation

Anirban Roy and Sinisa Todorovic
Oregon State University
Corvallis, OR 97330, USA
royani@oregonstate.edu, sinisa@eecs.oregonstate.edu

Abstract

This paper addresses the problem of weakly supervised semantic image segmentation. Our goal is to label every pixel in a new image, given only image-level object labels associated with training images. Our problem statement differs from common semantic segmentation, where pixel-wise annotations are typically assumed available in training. We specify a novel deep architecture which fuses three distinct computation processes toward semantic segmentation – namely, (i) the bottom-up computation of neural activations in a CNN for the image-level prediction of object classes; (ii) the top-down estimation of conditional likelihoods of the CNN's activations given the predicted objects, resulting in probabilistic attention maps per object class; and (iii) the lateral attention-message passing from neighboring neurons at the same CNN layer. The fusion of (i)-(iii) is realized via a conditional random field as recurrent network aimed at generating a smooth and boundary-preserving segmentation. Unlike existing work, we formulate a unified end-to-end learning of all components of our deep architecture. Evaluation on the benchmark PASCAL VOC 2012 dataset demonstrates that we outperform reasonable weakly supervised baselines and state-of-the-art approaches.

1. Introduction

This paper addresses the problem of semantic image segmentation under weak supervision. Given an image, our goal is to assign an object class label to every pixel. Knowledge about the objects is learned from training images with only image-level class labels, i.e., image tags. Our problem differs from fully supervised semantic segmentation, commonly addressed in previous work, where pixel-wise ground-truth annotations of object classes are available in training.

Semantic image segmentation is challenging as objects in the image may appear in various poses, under partial occlusion, and against cluttered background. This is a long-standing problem, addressed by a large number of successful approaches under the assumption of having access to ground-truth pixel labels in training [22, 14, 27, 11, 3, 7, 12]. Due to this assumption, it is difficult to extend previous work to a wide range of other domains that do not provide pixel-wise annotations, or provide an insufficient amount of such supervision for robust learning.

Toward relaxing the level of supervision required in training, recently, weakly supervised convolutional neural networks (CNNs) have been proposed for semantic image segmentation [24, 25, 28, 26, 45, 29, 18, 41, 4, 30, 36]. These approaches use only image tags in training. Most of them perform segmentation within the multi-instance learning (MIL) framework, which ensures that pixel labeling is consistent with predicting image tags, since the latter prediction can be readily used for specifying loss against the available image-level ground truth, and in this way train the CNN.

Inspired by the success of these approaches, we also start off with a CNN aimed at two tasks: pixel labeling and predicting image classes – where image classification results on training data are used for an end-to-end MIL-based learning. Specifically, we use the *DeepLab* net [7] for pixel labeling, and another fully-connected layer for predicting image classes. We then extend this framework so as to fuse top-down, bottom-up, and smoothness visual cues toward more accurate semantic segmentation, as illustrated in Fig. 1. Our extensions are aimed at addressing the following two issues that we have observed in segmentation results of related work [24, 25, 28, 26, 45, 29, 18]: (1) Poor localization of objects; and (2) Limited preservation of object boundaries and smoothness over the true spatial extents of objects.

To generate boundary preserving segmentation, we pass the pixel labels predicted by our CNN, along with raw pixels of the image, to a fully-connected conditional random field (CRF). Specifically, following [46], we implement the CRF as a recurrent neural network (RNN), and call this network CRF-RNN. Our CRF-RNN refines the initial CNN's prediction such that the pixel labels better fit image edges present in input image. Importantly, our CRF-based refinement of

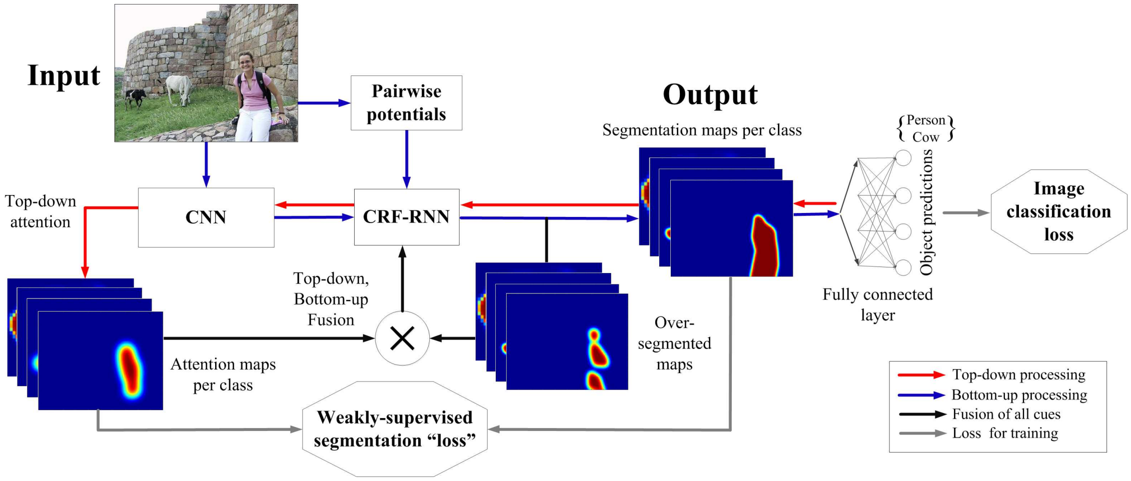


Figure 1: Overview: Given an image, we use a CNN to compute bottom-up segmentation maps for every object class (blue links for bottom-up computation). These pixel-wise predictions are aggregated with a fully-connected layer (FCL) for object recognition. The same CNN is used for top-down estimation of the attention maps for every recognized object class (red links for top-down computation). Finally, the bottom-up and top-down cues are fused and iteratively refined in the CRF-RNN for improving localization of object boundaries and spatial smoothness of the final segmentation (black links for fusion and refinement computation). In learning, we backpropagate the image classification loss, estimated from the FCL's outputs on on training images. This learning is regularized by the weakly supervised segmentation "loss", which estimates a distance between two probability distributions of pixel labels in the attention map and bottom-up segmentation (gray links for end-to-end training of all components).

segmentation is not an independent post-processing step, as in related work [24, 25, 29, 18], but an integral component of our deep architecture trained end-to-end. As the CRF-RNN tries to respect image edges, its output, however, may be prone to oversegmentation.

To avoid oversegmentation, and thus improve the smoothness over spatial extents of objects, we use our CNN for yet another, third task, that of predicting top-down visual attention maps of the recognized image classes. We specify the visual attention map of an object class as a spatial rectified Gaussian distribution [34, 16] of neural activations in the CNN for that class. This extends the recent approach [44] that uses a Markov chain to model parent-child dependencies of neural activations for estimating the attention map, since we estimate the rectified Gaussian distribution by accounting for three types of neural dependences in the CNN: (i) parent-to-child; (ii) child-to-parent; and (iii) between activations of neighboring neurons at the same CNN layer. Importantly, we compute the attention map using the same CNN aimed at semantic segmentation, unlike related work that uses an external network for estimating object seeds [18].

As shown in Fig. 1, our approach iteratively: (i) Fuses the oversegmentation maps produced bottom-up by the CNN and CRF-RNN, and the attention maps estimated top-down by the CNN, and then (ii) Refines the fused pixel-label predictions with the CRF-RNN to generate the final segmentation maps. The attention map represents discriminative object parts critical for classification, whereas the segmentation map captures the object's spatial extent. We adopt the same definition of top-down and bottom-up processing used in the related literature [5, 20, 16] where the bottom-up process predicts object classes from pixels, and the top-down process predicts the attention map in the image conditioned on the object class predicted by the bottom-up process. All components of our deep architecture are trained end-to-end by estimating the image classification loss and segmentation "loss". Image classes are predicted in training using a fully-connected layer (FCL) from the pixel label predictions. This, in turn, generates the image classification loss that is backpropagated through the FCL, CRF-RNN, and CNN for learning all network parameters. Regrading segmentation "loss", we here slightly abuse the common definition of loss, since we are not given ground-truth segmentation. We estimate segmentation "loss" as a distance between the two probability distributions of pixel labels in the attention map and bottom-up segmentation. The segmentation "loss" is backpropagated through the CRF-RNN and CNN and serves to regularize the image classification loss.

Evaluation on the benchmark PASCAL VOC 2012 dataset demonstrates that we outperform reasonable weakly supervised baselines and state-of-the-art approaches.

Our contributions include:

- New deep architecture that fuses top-down attention and bottom-up segmentation, and refines segmentation for preserving boundaries. The architecture is unified, and does not use external networks, nor post-processing.
- New modeling of the visual attention map using the rectified Gaussian distribution which accounts for statistical dependencies between activations of parents, children, and neighboring neurons in the CNN.

In the following, Sec. 2 reviews related work, Sec. 3 specifies our bottom-up pixel labeling and aggregation for object recognition, Sec. 4 formulates our top-down attention estimation, Sec. 5 describes our boundary-preserving refinement of segmentation, Sec. 6 explains the two loss functions and our learning, and Sec. 7 presents our results.

2. Prior Work

Weakly supervised semantic image segmentation has been addressed using graphical models, and parametric structured prediction models [39, 40, 9, 45, 21]. These approaches typically exploit heuristics about spatial smoothness (e.g., based on similarity between neighboring pixels [39]), require pre-processing for extracting superpixels [40], or use weak segmentation priors [9]. Recently, CNN-based methods [28, 25, 24, 4, 41] are shown to achieve better performance by typically considering multiple-instance learning (MIL) for iteratively reinforcing that their output segmentations are consistent with ground-truth image tags. The MIL framework can be extended with generalized expectation or posterior regularization for maximizing the expectation of model parameters under domain constraints [25].

Toward improving performance, some recent approaches [28, 24, 4, 41] seek to initialize object localization by running detectors of object proposals [1, 10]. However, this increases the level of supervision, since object-proposal detectors require bounding-box annotations or object boundary annotation for training. Also, using attention-based object localization has been shown to improve weakly supervised segmentation [29, 18, 15, 30]. However, these approaches typically resort to external networks for computing attention cues [18], or estimate foreground masks (class non-specific) from neural activations bottom-up [29]. Recent work [44] computes attention maps by estimating a top-down Markov chain, but this work does not consider weakly supervised segmentation. We extend the Markov chain formulation of [44] by using the rectified Gaussian distribution for modeling of the visual attention map, resulting in an improved spatial smoothness of our attention maps per object class.

Combining top-down and bottom-up cues for image segmentation and other vision problems is a recurring research topic; however, the two cues are often computed in separate stages [5, 6, 20, 42]. Recent approach [16] combines the two computation processes in a single CNN for human pose estimation using the rectified Gaussian distribution. But their CNN is trained under full supervision. While we address a different vision problem, the key differences are that our covariance matrix of the rectified Gaussian is not binary at theirs but estimated based on visual appearance, and our top-down attention cues are semantically meaningful conditioned on the predicted object classes.

3. Bottom-Up Computation Process

Pixel labeling. Given an image, $\boldsymbol{x}$, we use the *DeepLab* net with large field of view [7] to generate pixel labels $\boldsymbol{y} = \{y_i\}$, where $y_i \in \mathcal{Y}$ is the object class label of ith pixel from the set of object classes $\mathcal{Y}$. Specifically, we generate $K = |\mathcal{Y}|$ segmentation maps, by computing output score $f_i(y)$ at every pixel i for every object class $y \in \mathcal{Y}$. The pixel-wise scores are normalized to estimate the corresponding posteriors using the standard soft-max operation as

$$p_i^S(y|\boldsymbol{x}) = \frac{e^{f_i(y)}}{\sum_{y\in\mathcal{Y}} e^{f_i(y)}}, \tag{1}$$

Henceforth, we will use the shorthand notation $p_i^S(y)$ to denote $p_i^S(y|\boldsymbol{x})$ as the bottom-up segmentation prediction.

Aggregation. The above pixel-wise object prediction scores are then aggregated for object recognition, i.e., predicting the set of object classes, $\mathcal{Y}_{\boldsymbol{x}}$, that are present in image $\boldsymbol{x}$. The literature presents a host of heuristic methods for such aggregation, including global max pooling (GMP) [23], global average pooling (GAP) [47], log-sum-exponential (LSE) measure for a smooth combination of GMP and GAP [28], and global weighted rank pooling (GWRP) to favor high scores for ground-truth objects and suppressing others in the aggregation [18]. Instead of using these heuristic methods, we employ a fully connected layer (FCL) to estimate the image-level scores from the pixel-wise scores, and then train the FLC together with other components of our approach. Given K maps of pixel-wise scores $\{f_i(y) : i, y \in \mathcal{Y}\}$, the FLC outputs K normalized object scores, $\{p(y|\boldsymbol{x}) : y \in \mathcal{Y}\}$. For this, each segmentation heatmap $\{f_i(y) : i\}$ is fully connected to the corresponding output unit representing the object class y.

4. Top-Down Computation Process

This section explains how to estimate top-down visual attention maps for every object class, which are then used as contextual cues for improving object localization and reducing oversegmentation in the bottom-up pixel labeling. Following a long line of work on estimating probabilistic visual attention maps [38, 17, 35], as well as recent approaches to visualizing neural activations [43, 32, 2, 44], we use a top-down computation process for estimating probabilistic visual attention maps of neural activations at each layer of our CNN. Our top-down estimation is performed one layer at a time, starting from the FCL's output layer for object recognition, described in Sec. 3. For efficiency, as in [44], we stop our top-down computation at the `pool-4` layer, and then upscale this result to the image size for obtaining the K probabilistic visual attention maps, over all pixels, corresponding to K object classes.

We define a visual attention of ith neuron at layer l, for object class y, as a relevance of the neuron's activation, a_i^l, for predicting y in the image – denoted as $p(a_i^l|y) \geq 0$. The visual attention map of all neurons at layer l is defined as a vector of random variables: $\boldsymbol{p}^l(y) = [\dots p(a_i^l|y) \dots]^\top$, governed by the rectified Gaussian distribution [34, 16]:

$$P(\boldsymbol{p}) \propto \exp(\frac{1}{2}\boldsymbol{p}^\top D\boldsymbol{p} + \boldsymbol{b}^\top\boldsymbol{p}), \quad \boldsymbol{p} \succeq \boldsymbol{0} \tag{2}$$

where we use the shorthand notation $\boldsymbol{p} = \boldsymbol{p}^l(y)$, matrix $D = D^l = [\delta_{ii'}^l]$ captures the strength of dependencies between neighboring neural activations at the same layer l, and $\boldsymbol{b} = \boldsymbol{b}^l(y)$ represents parent-child dependences of neural activations at layer l and next layer $l-1$. By design, we guarantee $\delta_{ii'}^l < 0$, and thus the negative $-D$ is a copositive matrix in (2), i.e., $-\boldsymbol{p}^\top D\boldsymbol{p} \geq 0$, for $\boldsymbol{p} \succeq \boldsymbol{0}$. Computation of $\delta_{ii'}^l$ is explained below.

From (2), it follows that computation of $\boldsymbol{p}^l(y)$ amounts to the MAP estimation of the rectified Gaussian, which in turn can be formulated as the quadratic program with non-negativity constraints:

$$\max_{\boldsymbol{p}\succeq\boldsymbol{0}} \frac{1}{2}\boldsymbol{p}^T D\boldsymbol{p} + \boldsymbol{b}^\top\boldsymbol{p}. \tag{3}$$

The copositive property the negative matrix $-D$ guarantees convergence of the quadratic optimization in (3) [16, 34].

Given our CNN, we sequentially compute the quadratic program in (3) top-down, one layer at a time, until the `pool-4` layer. The results from previous layer $l-1$ are used to define the parameters $\boldsymbol{b}^l(y)$ as they capture parent-child dependences of neural activations. Finally, the estimated $\boldsymbol{p}^l(y)$ for the `pool-4` layer is upscaled to the image size, and then normalized over the object classes in order to make a proper probability distribution at every pixel:

$$p_i^A(y) = \frac{p(a_i^l|y)}{\sum_{y\in\mathcal{Y}} p(a_i^l|y)}. \tag{4}$$

In comparison with the recent work [44], which accounts only for parent-child dependences of neural activations as $p(a_i^l|y) = \sum_j p(a_i^l|a_j^{l-1})p(a_j^{l-1}|y)$, we increase computation time by a small margin, but considerably improve the attention maps such that they cover well the true spatial extents of objects.

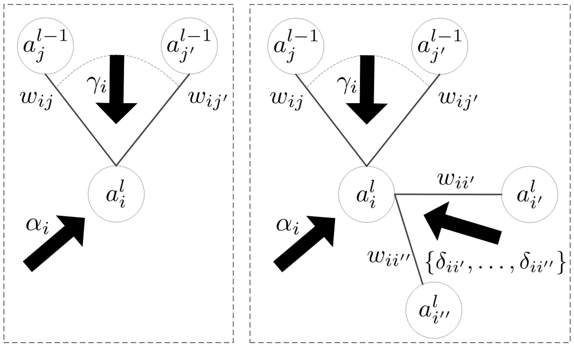
In the rest of this section, we define the parameters $\boldsymbol{b}^l(y)$ and D of the rectified Gaussian. As illustrated in Fig. 2, each $p(a_i^l|y)$ is made dependent on the neural activations of:

1. parents: $\gamma_i^l(y)$;
2. feed-forward neural processing: α_i^l;
3. neighboring neurons at the same layer l, $D^l = [\delta_{ii'}^l]$.

where we compute $\boldsymbol{b}^l(y) = \gamma_i^l(y) + \alpha_i^l$.

First, following [44], we define dependence of $p(a_i^l|y)$ on the activations of parent neurons $\mathcal{P}_i$ in the CNN as

$$\gamma_i^l(y) = \sum_{j\in\mathcal{P}_i} p(a_i^l|a_j^{l-1})p(a_j^l|y), \tag{5}$$



Prior work

Our formulation

Figure 2: (Left) Previous work [44] computes a top-down Markov chain for estimating the attention map. (Right) We additionally consider neighboring neural activations at the same layer for improving estimation of the attention map based on smoothness of objects.

where $p(a_i^l|a_j^{l-1})$ is the transition probability defined as

$$p(a_i^l|a_j^{l-1}) = \frac{w_{ij}^+ \cdot a_i^l}{\sum_{i'\in\mathcal{C}_j} w_{i'j}^+ \cdot a_{i'}^l}, \tag{6}$$

where $w_{ij}^+ = \max\{0, w_{ij}\}$ accounts only for positive weights between neurons i and their parents j in the CNN, and $\mathcal{C}_j$ denotes the set of children of j.

Second, we normalize the feed-forward neural processing in the CNN at neuron i across all neurons at the same layer l, resulting in

$$\alpha_i^l = \frac{a_i^l}{\sum_{i'} a_{i'}^l} \tag{7}$$

Finally, third, we use the standard bi-lateral filtering of pixels in the image to define the strength $D^l = [\delta_{ii'}^l]$ of dependencies between neighboring neurons at the same layer. For every neuron pair (i, i') at layer l, we determine their corresponding centers of the pixel areas in the image that the neurons have access to, and compute their bi-lateral similarity [37] as

$$w_{ii'}^l = \exp(-\frac{\| \boldsymbol{z}_i - \boldsymbol{z}_{i'} \|^2}{\sigma_z^2})\ \exp(-\frac{\| \boldsymbol{r}_i - \boldsymbol{r}_{i'} \|^2}{\sigma_r^2}), \tag{8}$$

where $\boldsymbol{z}_i = (\mathrm{x}_i, \mathrm{y}_i)$ is the pixel location, and $\boldsymbol{r}_i$ is the HSV color histogram, and $\sigma_z = 10$ and $\sigma_r = 30$ control sensitivity. Then, we normalize the bi-lateral similarity, and define

$$\delta_{ii'}^l = \frac{w_{ii'}}{\sum_{i'} w_{ii''}} - 1, \tag{9}$$

Note that $\delta_{ii'}^l$ depends only on the image, and not the object class prediction, and hence can be pre-computed for efficiency. Note that in (8), $w_{ii} = 1$ and $w_{ii'} > 0 : i \neq i'$ which implies $\delta_{ii'} < 0$ in (9). Thus, the matrix $-D$ is copositive which guarantees the convergence of the quadratic optimization in (3). Recall that the attention maps are used to compute the segmentation loss which is backpropagated through the network during learning (Fig. 1).

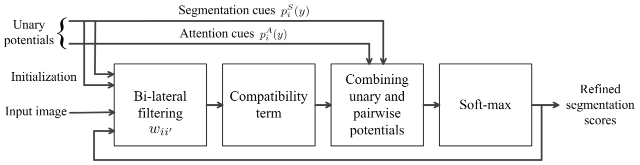

Segmentation cues $p_i^S(y)$
Unary potentials
Attention cues $p_i^A(y)$
Initialization
Input image
Bi-lateral filtering $w_{ii'}$
Compatibility term
Combining unary and pairwise potentials
Soft-max
Refined segmentation scores

Figure 3: A single step of the CRF mean-field inference implemented as a stack of convolution layers. The mean-field iterations represent a recurrent neural network.

5. Refinement Computation Process

Due to the successive pooling in the CNN, the initial pixel labeling $p_i^S(y|\boldsymbol{x})$, given by (1), is likely to produce a coarse segmentation map, with poor detections of object boundaries. To address this issue, we pass the initial coarse segmentation, together with the estimated visual attention maps and the input image, to the CRF-RNN [19, 46], for refining the segmentation. We consider a fully connected CRF, whose energy of pixel-wise class assignment $\boldsymbol{y} = \{y_i\}$ is defined as

$$E(\boldsymbol{y}) = \sum_i \phi(y_i) + \sum_{(i,i')} \psi(y_i, y_{i'}), \tag{10}$$

where $\phi(y_i)$ and $\psi(y_i, y_{i'})$ denote the unary and pairwise potentials, specified below.

Unary Potential. To initialize our segmentation refinement, at every pixel i, we combine the top-down visual attention $p_i^A(y_i)$, given by (4), and the bottom-up segmentation $p_i^S(y|\boldsymbol{x})$, given by (1), for computing the corresponding unary potential $\phi(y_i)$ as

$$\phi(y_i) = -\log(p_i^S(y_i)p_i^A(y_i)). \tag{11}$$

In (11), we use the visual attention $p_i^A(y_i)$ as a pixel-level prior (PLP) for segmentation. Our motivation comes from existing work on weakly supervised segmentation that has considered image level priors (ILP), such as, e.g., image-level object prediction score, for refining their segmentation [28, 31, 39]. ILPs have been shown to improve weakly supervised segmentation by reducing false positives. Our formulation in (11), extends this work, because our attention driven prior at every pixel seems more suitable for segmentation than the image-level prior. Unlike ILP, our PLP incorporates cues about object's location.

Pairwise Potential. We define the pairwise potential in terms of the bi-lateral weights $w_{ii'}$, given by (8), for ensuring that our segmentation refinement respects object boundaries:

$$\psi(y_i, y_{i'}) = \mu(y_i, y_{i'})w_{ii'}, \tag{12}$$

where $\mu(y_i, y_{i'})$ is the label compatibility aimed at estimating the likelihood of co-occurrence of classes y_i and $y_{i'}$ at pixels i and i'. Note that $\mu(y_i, y_{i'})$ varies over different pixel locations. It is implemented as a convolutional layer, and learned from the segmentation "loss" specified in Sec. 6.

CRF Inference. Following [19, 46], we conduct CRF inference as a series of mean-field iterations. As shown in Fig. 3, each mean-field estimation corresponds to the feed-forward neural processing along a stack of convolutional layers, the result of which is fed back for another iteration. Hence, the mean-field iterations represent a recurrent neural network. Note that, our CRF inference takes $\boldsymbol{p}^S$ and $\boldsymbol{p}^A$ as inputs to compute $\phi(y_i)$, as in (11), for every pixel i, and all K object classes. In the first iteration, soft-max scores over the unary potentials are considered as marginal probabilities to initialize the solution. In the following iterations, the marginal probabilities are estimated as soft-max scores of the CRF-RNN output. Bi-lateral filter responses are computed from the input image given the output of previous layer. Unlike [46], we consider fixed bi-lateral kernels as they cannot be reliably learned without pixel-wise supervision. The label compatibility $\mu(y_i, y_{i'})$ is estimated by applying 1×1 convolution filters with K input and K output channels for K object classes, given the bi-lateral responses. Finally, given the estimated $\phi(y_i)$ and $\psi(y_i, y_{i'})$, the combined CRF potentials are passed through the soft-max operation for generating the normalized segmentation scores for the next CRF-RNN iteration.

6. End-to-End Learning and Loss Functions

All components of our approach are learned in an end-to-end manner, using only ground-truth image tags. In order to use this image-level supervision in learning, our approach aggregates the predicted pixel labels on training images into object recognition, which in turn can be used to estimate the classification loss Δ^C. For training the CRF-RNN and our initial segmenter *DeepLab* network [7], we additionally use an object segmentation "loss", Δ^S, defined relative to the visual attention map estimated on training images, since we do not have access to pixel-wise annotations. Thus, in our learning, we backpropagate the following loss:

$$\Delta = \Delta^C + \lambda\Delta^S, \tag{13}$$

where $\lambda = 1.5$ is set by cross validation.

Classification Loss is defined in terms of the FCL's output aggregation function for the image-level object recognition, $p(y|\boldsymbol{x})$, specified in Sec. 3 as

$$\Delta^C = -\frac{1}{|\mathcal{Y}_{\boldsymbol{x}}|}\sum_{y\in\mathcal{Y}_{\boldsymbol{x}}} \log\, p(y|\boldsymbol{x}) - \frac{1}{|\bar{\mathcal{Y}}_{\boldsymbol{x}}|}\sum_{y\in\bar{\mathcal{Y}}_{\boldsymbol{x}}} \log\, (1-p(y|\boldsymbol{x})) \tag{14}$$

where $\mathcal{Y}_{\boldsymbol{x}}$ denotes a set of ground-truth object classes present in training image $\boldsymbol{x}$, and $\bar{\mathcal{Y}}_{\boldsymbol{x}} = \mathcal{Y} \setminus \mathcal{Y}_{\boldsymbol{x}}$ is a set of classes that are known to be absent. Δ^C penalizes low prediction scores from the FCL for the objects annotated as present in the training image, and high scores for the other objects.

Object Segmentation Loss is defined to penalize any discrepancies between the estimated segmentation and visual attention maps. Δ^S is defined as a distance between the two predicted distributions $\boldsymbol{p}^S$ and $\boldsymbol{p}^A$ for the objects $y \in \mathcal{Y}_{\boldsymbol{x}}$ annotated as present in the training image:

$$\Delta^S = -\frac{1}{N \cdot |\mathcal{Y}_{\boldsymbol{x}}|} \sum_{i=1}^{N} \sum_{y \in \mathcal{Y}_{\boldsymbol{x}}} [p_i^A(y) \ \log p_i^S(y) + (1 - p_i^A(y)) \ \log(1 - p_i^S(y))], \tag{15}$$

where N denotes the number of pixels.

It is worth noting that the visual attention is used in two different ways in our approach – namely, for computing the unary potentials of the CRF-RNN inference on a single image, and for estimating Δ^S in learning on a mini-batch of training images. Hence, these two uses of the visual attention in our approach are not redundant, as also demonstrated in our experiments.

7. Experiments

In this section, we first describe our experimental setup and then present the results.

Dataset. We evaluate our approach on the PASCAL VOC 2012 dataset [13] which is commonly considered as the weakly supervised segmentation benchmark [25, 28, 29, 18]. This dataset consists of 21 object classes including the background. We follow the standard experimental setup where the images are split into three sets: 1464 training images, 1456 test images, and 1449 validation images. Following common practice, we consider additional *trainaug* set in training [18, 36] and evaluate our approach on images in validation and test sets. We consider the standard PASCAL VOC segmentation metric which is defined as mean intersection over union (mIoU) ratio, also known as Jaccard index.

Implementation details. We consider the *DeepLab* network with large field of view [7] for image segmentation. *DeepLab* adopts the VGG-16 net [33] for segmentation by replacing fully connected layers with convolutional layers. Given an input image, *DeepLab* produces coarse heatmaps corresponding to each object class. The network is trained using sub gradient descent with momentum. We consider a batch size of 20 images and the momentum is set to 0.9. The learning rate is initially set to 0.001 and decreased by a factor of 10 in every 2000 iterations. We train the network for 10000 iterations. Overall training takes $\approx$ 10 hours on an Nvidia Tesla k80 GPU, which is comparable to [25, 24]. During inference, we first compute object-specific attention maps which are then considered as the attention based PLP. The attention based PLP serves as the unary potentials in the CRF-RNN layer. Though image-level tags are not available during inference, attention maps can be estimated from reliable object predictions based on *fully supervised* image classification. For both learning and inference, we apply CRF-RNN for three iterations as additional iterations do not have a significant effect on the final performance.

Baselines. To justify the importance of various components of our approach, we define the following baselines. Comparisons with the baselines are shown in Tab. 1.

B1: Without top-down attention (w/o att). In this baseline, we ignore the top-down attention cues in our approach. As the segmentation loss cannot be computed without attention cues, only the classification loss is used to learn the segmentation network. Note that without attention cues, localizing objects in the image is difficult. Results in Tab. 1 show that ignoring top-down attention cues has a significant effect on the performance which justifies the importance of attention cues in weakly supervised segmentation.

B2: Without segmentation-loss (w/o seg loss). In this baseline, we ignore the segmentation loss which is computed based on the attention maps. Without segmentation loss, attention cues are considered only in unary potentials in the CRF-RNN layer as defined in (11). As shown in Tab. 1 that the segmentation loss is important for weakly supervised segmentation as it is required in the learning the segmentation net and CRF-RNN.

B3: Without attention based unary potential (w/o att unary). In this baseline, we do not consider attention cues in unary potentials in the CRF-RNN layer. Thus, the attention cues are incorporated in the segmentation framework only through the loss function (15). The results in Tab. 1 show that considering attention cues in CRF unary potentials improves overall performance.

B4: Without considering neighboring dependences in attention (w/o neighbor). In this baseline, we ignore the dependences of neighboring neurons (i.e., $\delta_{ii'}^l$ in (9)) while computing the attention maps. Thus, we only consider parent-child dependences to compute attention maps as in [44]. As shown in Tab. 1, performance is inferior without considering the neighboring dependences in the attention estimation as these dependences provides the cues about object's smoothness and boundary.

B5: Without CRF-RNN layer (w/o CRF-RNN). In this baseline, instead of applying the CRF-RNN layer to refine the segmentation maps, a dense CRF based post-processing is performed [7]. Without the CRF-RNN layer, the label compatibility or the co-occurrence between the object classes (i.e., $\mu(y_i, y_{i'})$ in (12)) cannot be learned. We see in Tab. 1 that considering CRF-RNN layer achieves better performance than CRF based post-processing.

B6: Without attention cues in inference (w/o att inference). In this baseline, we ignore the attention cues to compute CRF unary potentials during inference. Recall that, attention cues can be considered as the pixel-level priors which are important to localize objects in the image. Thus, ignoring attention in inference results in worse overall performance (Tab. 1).

Comparisons with the state-of-the-art with image-

	background	aeroplane	bike	bird	boat	bottle	bus	car	cat	chair	cow	diningtable	dog	horse	motorbike	person	plant	sheep	sofa	train	tv/monitor	avg.
MIL+ILP [28]	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
EM [24]	67.2	29.2	17.6	28.6	22.2	29.6	47.0	44.0	44.2	14.6	35.1	24.9	41.0	38.8	41.6	32.1	24.8	37.4	24.0	38.1	31.6	33.8
CCNN [25]	68.5	25.5	18.0	25.4	20.2	36.3	46.8	47.1	48.0	15.8	37.9	21.0	44.5	34.5	46.2	40.7	30.4	36.4	22.2	38.8	36.9	35.6
DSCM [30]	76.7	45.1	24.6	40.8	23.0	34.8	61.0	51.9	52.4	15.5	45.9	32.7	54.9	48.6	57.4	51.8	**38.2**	55.4	32.2	42.6	39.6	44.1
F-B [29]	79.2	60.1	20.4	50.7	**41.2**	**46.3**	62.6	49.2	62.3	13.3	49.7	**38.1**	58.4	49.0	57.0	48.2	27.8	55.1	29.6	**54.6**	26.6	46.6
SEC [18]	82.4	62.9	26.4	61.6	27.6	38.1	66.6	62.7	75.2	**22.1**	53.5	28.3	65.8	57.8	62.3	52.5	32.2	62.6	32.1	45.4	45.3	50.7
Our approach	**85.8**	**65.2**	**29.4**	**63.8**	31.2	37.2	**69.6**	**64.3**	**76.2**	21.4	**56.3**	29.8	**68.2**	**60.6**	**66.2**	**55.8**	30.8	**66.1**	**34.9**	48.8	**47.1**	**52.8**

Table 2: Comparison with the state-of-the-art approaches on PASCAL 2012 validation set in terms of mIOU measure (%).

	PASCAL validation	PASCAL test
w/o att	30.5	31.6
w/o seg loss	47.5	49.1
w/o att unary	50.1	51.4
w/o neighbor	51.3	52.1
w/o CRF-RNN	49.4	51.3
w/o att inference	50.4	51.8
Full approach	**52.8**	**53.7**

Table 1: Comparisons with the baseline approaches on PASCAL 2012 validation and test datasets in terms of mIoU measure (%).

	val.	test	Additional supervision
MIL+ILP+SP-bb [28]	37.8	37.0	BING bounding box
MIL+ILP+SP-seg [28]	42.0	40.6	MCG object proposals
SN-B [41]	41.9	43.2	MCG object proposals
EM-Adapt+crop [24]	38.2	39.6	Multiple image crops
CCCN+crop[25]	36.4	47.2	Multiple image crops
CCCN+size [25]	42.4	-	Object size (big or small)
Point click [4]	43.4	-	1 click per object instance
Check mask [29]	**51.5**	**52.9**	User selected fore-ground mask

Table 4: Comparisons among the approaches that use addition supervision on PASCAL 2012 validation and test sets in terms of mIoU measure (%).

	PASCAL validation	PASCAL test
GMP	48.3	49.6
GAP	47.4	48.3
LSE	50.8	51.3
Our FCL	**52.8**	**53.7**

Table 5: Comparisons of FCL with other aggregation methods in terms of mean IoU measure (%) on PASCAL 2012 validation and test datasets.

level annotation. Comparisons with the state-of-the-art approaches are performed on PASCAL 2012 validation and test images. Our approach is learned with only image-level tags. Thus, for fair comparison, we compare with the approaches which consider *only* image level annotations as weak supervision. Due to the attention based localization cues, we do not need to rely on additional supervisions such as object proposals [28], image crops [24] or size of the objects [25]. The results on PASCAL validation set and test set are shown in Tab. 2 and Tab. 3, respectively, where we outperform the state-of-the-art approaches in terms of mIoU metric.

Comparisons with the state-of-the-art with additional annotation. Some approaches consider additional low-cost supervision to facilitate weakly supervised segmentation. For example, MIL+ILP+SP-bb [28], MIL+ILP+SP-seg [28], SN-B [41] use object localization cues in terms on MCG object proposals [1] or BING bounding boxes [10]. Variant of EM-Adapt [24] and CCCN [25] consider multiple image crops to increase the amount of supervision. Additional 'click per object' annotations are used in [4] and CCCN+size [25] consider additional 1-bit supervision in terms of size (big or small) of an object. It is unfair to directly compare with these approach, but we summarize the results of the above-mentioned approaches in Tab. 4 for completeness.

Evaluation of the aggregation methods. Recall that we consider a fully connected layer (FCL) which is learned to aggregate pixel-wise predictions into a image-level object prediction score. We compare FCL with other aggregation methods such as GMP [23], GAP [47], and LSE [28] which is a smooth combination of GMP and GAP. Though considering the FCL layer increases the number of parameters in learning, as shown in Tab. 5, our proposed FCL significantly outperforms other heuristic based aggregation methods.

Qualitative results. In Fig. 4, we present the qualitative results on the PASCAL 2012 validation set. Our approach, by most part, can correctly localize objects in images and respects the object boundaries. A pair of failure cases are shown in Fig. 5, where our approach fails to detect the 'aeroplane' in an image due to its uncommon appearance with respect to other 'aeroplane' instances in training data. In the second case, our approach fails to detect a few 'person' instances and 'bottles' in the image. We believe this is due to missing attention cues for the small objects (e.g., bottle) in the image. Note that segmenting small objects is challenging even with full supervision [8].

8. Conclusion

We have specified a new deep architecture for weakly supervised image segmentation. Our key idea is to estimate and fuse bottom-up, top-down, and smoothness cues using the

	background	aeroplane	bike	bird	boat	bottle	bus	car	cat	chair	cow	diningtable	dog	horse	motorbike	person	plant	sheep	sofa	train	tv/monitor	avg.
MIL [26]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	25.66
MIL+ILP [28]	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
CCNN [25]	-	24.2	19.9	26.3	18.6	38.1	51.7	42.9	48.2	15.6	37.2	18.3	43.0	38.2	52.2	40.0	33.8	36.0	21.6	33.4	38.3	35.6
DSCM [30]	78.1	43.8	26.3	49.8	19.5	40.3	61.6	53.9	52.7	13.7	47.3	34.8	50.3	48.9	69.0	49.7	**38.4**	57.1	34.0	38.0	40.0	45.1
F-B [29]	80.3	57.5	24.1	66.9	**31.7**	43.0	67.5	48.6	56.7	12.6	50.9	42.6	59.4	52.9	65.0	44.8	41.3	51.1	33.7	**44.4**	33.2	48.0
SEC [18]	83.5	56.4	28.5	64.1	23.6	**46.5**	70.6	58.5	71.3	**23.2**	54.0	**28.0**	68.1	62.1	70.0	55.0	**38.4**	58.0	39.9	38.4	48.3	51.7
Our approach	**85.7**	**58.8**	**30.5**	**67.6**	24.7	44.7	**74.8**	**61.8**	**73.7**	22.9	**57.4**	27.5	**71.3**	**64.8**	**72.4**	**57.3**	37.0	**60.4**	**42.8**	42.2	**50.6**	**53.7**

Table 3: Comparison with the state-of-the-art approaches on PASCAL 2012 test set in terms of mIOU measure (%).

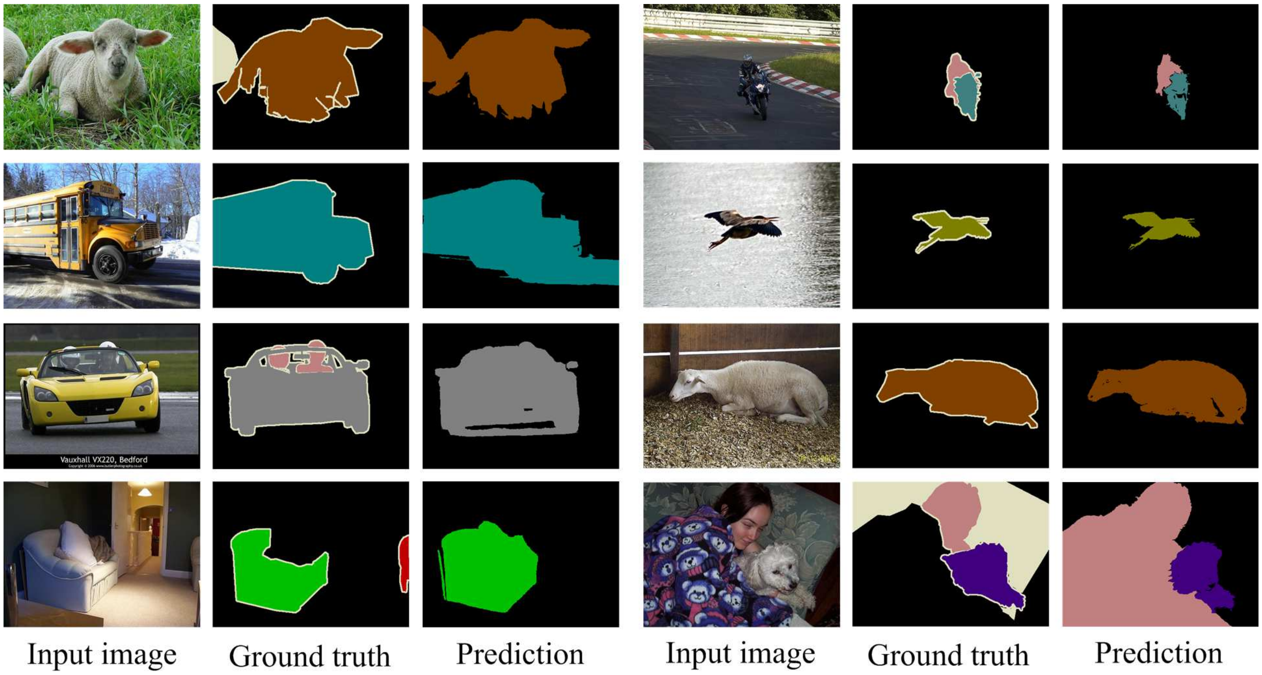


Figure 4: Qualitative results on PASCAL 2012 validation set.

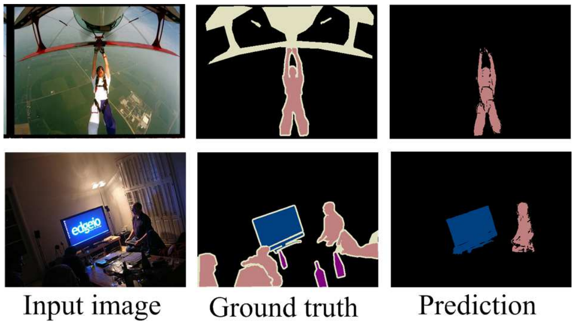


Figure 5: Failure cases on PASCAL 2012 validation set.

same network toward better fitting to object boundaries and covering spatial extents of objects. Our unified framework consists of a CNN, CRF-RNN, and a fully connected layer, which can be trained end-to-end using only ground-truth image tags. In our evaluation on the benchmark PASCAL VOC 2012 dataset, we have observed that our approach can localize objects without having to rely on additional supervision such as object proposals and image crops. Estimating appearance based neighboring dependencies for visual attention enabled us to better localize the full extent of objects rather than just parts. A comparison of our approach with the baselines has justified the importance of attention cues, CRF-RNN smoothing, and FCL layer as an aggregation method in weakly supervised segmentation.

Acknowledgment

This work was supported in part by DARPA XAI and NSF RI1302700.

References

[1] P. Arbeláez, J. Pont-Tuset, J. T. Barron, F. Marques, and J. Malik. Multiscale combinatorial grouping. In *CVPR*, 2014. 3, 7

[2] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7), 2015. 3

[3] V. Badrinarayanan, A. Handa, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling. *arXiv preprint arXiv:1505.07293*, 2015. 1

[4] A. Bearman, O. Russakovsky, V. Ferrari, and L. Fei-Fei. What's the point: Semantic segmentation with point supervision. In *ECCV*, 2016. 1, 3, 7

[5] E. Borenstein and S. Ullman. Combined top-down/bottom-up segmentation. *PAMI*, 30(12):2109–2125, 2008. 2, 3

[6] T. Brox, L. Bourdev, S. Maji, and J. Malik. Object segmentation by alignment of poselet activations to image contours. In *CVPR*, 2011. 3

[7] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected CRFs. In *ICLR*, 2015. 1, 3, 5, 6

[8] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille. Attention to scale: Scale-aware semantic image segmentation. In *CVPR*, 2015. 7

[9] X. Chen, A. Shrivastava, and A. Gupta. Enriching visual knowledge bases via object discovery and segmentation. In *CVPR*, 2014. 3

[10] M.-M. Cheng, Z. Zhang, W.-Y. Lin, and P. Torr. Bing: Binarized normed gradients for objectness estimation at 300fps. In *CVPR*, 2014. 3, 7

[11] C. Couprie, C. Farabet, L. Najman, and Y. LeCun. Indoor semantic segmentation using depth information. In *ICLR*, 2013. 1

[12] J. Deng, N. Ding, Y. Jia, A. Frome, K. Murphy, S. Bengio, Y. Li, H. Neven, and H. Adam. Large-scale object classification using label relation graphs. In *ECCV*. 2014. 1

[13] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html. 6

[14] C. Farabet, C. Couprie, L. Najman, and Y. LeCun. Learning hierarchical features for scene labeling. *PAMI*, 35(8):1915–1929, 2013. 1

[15] S. Hong, J. Oh, B. Han, and H. Lee. Learning transferrable knowledge for semantic segmentation with deep convolutional neural network. In *CVPR*, 2016. 3

[16] P. Hu, D. Ramanan, J. Jia, S. Wu, X. Wang, L. Cai, and J. Tang. Bottom-up and top-down reasoning with hierarchical rectified gaussians. In *CVPR*, 2016. 2, 3, 4

[17] C. Koch and S. Ullman. Shifts in selective visual attention: Towards the underlying neural circuitry. In *Matters of Intelligence*, pages 115–141. Springer, 1987. 3

[18] A. Kolesnikov and C. H. Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *ECCV*, 2016. 1, 2, 3, 6, 7, 8

[19] P. Krähenbühl and V. Koltun. Efficient inference in fully connected CRFs with gaussian edge potentials. In *NIPS*, 2011. 5

[20] M. P. Kumar, P. H. Torr, and A. Zisserman. Objcut: Efficient segmentation using top-down and bottom-up cues. *PAMI*, 32(3):530–545, 2010. 2, 3

[21] B. Lai and X. Gong. Saliency guided dictionary learning for weakly-supervised image parsing. In *CVPR*, 2016. 3

[22] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. *CVPR*, 2015. 1

[23] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Is object localization for free?-weakly-supervised learning with convolutional neural networks. In *CVPR*, 2015. 3, 7

[24] G. Papandreou, L.-C. Chen, K. P. Murphy, and A. L. Yuille. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *ICCV*, 2015. 1, 2, 3, 6, 7

[25] D. Pathak, P. Krahenbuhl, and T. Darrell. Constrained convolutional neural networks for weakly supervised segmentation. In *CVPR*, 2015. 1, 2, 3, 6, 7, 8

[26] D. Pathak, E. Shelhamer, J. Long, and T. Darrell. Fully convolutional multi-class multiple instance learning. In *ICLR*, 2015. 1, 8

[27] P. O. Pinheiro and R. Collobert. Recurrent convolutional neural networks for scene parsing. *ICML*, 2014. 1

[28] P. O. Pinheiro and R. Collobert. Weakly supervised semantic segmentation with convolutional networks. In *CVPR*, 2015. 1, 3, 5, 6, 7, 8

[29] F. Saleh, M. S. A. Akbarian, M. Salzmann, L. Petersson, S. Gould, and J. M. Alvarez. Built-in foreground/background prior for weakly-supervised semantic segmentation. In *ECCV*, 2016. 1, 2, 3, 6, 7, 8

[30] W. Shimoda and K. Yanai. Distinct class-specific saliency maps for weakly supervised semantic segmentation. In *ECCV*, 2016. 1, 3, 7, 8

[31] J. Shotton, M. Johnson, and R. Cipolla. Semantic texton forests for image categorization and segmentation. In *CVPR*, 2008. 5

[32] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *ICLR*, 2014. 3

[33] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 6

[34] N. D. Socci, D. D. Lee, and H. Sebastian Seung. The rectified gaussian distribution. 1998. 2, 4

[35] Y. Tang, N. Srivastava, and R. Salakhutdinov. Learning generative models with visual attention. In *NIPS*, 2014. 3

[36] P. Tokmakov, K. Alahari, and C. Schmid. Learning semantic segmentation with weakly-annotated videos. In *ECCV*, 2016. 1, 6

[37] C. Tomasi and R. Manduchi. Bilateral filtering for gray and color images. In *ICCV*, 1998. 4

[38] J. Tsotsos, S. Culhane, W. Wai, Y. Lai, N. Davis, and F. Nuflo. Modeling visual attention via selective tuning. *Artificial Intelligence*, 78(1):507–545, 1995. 3

[39] A. Vezhnevets and J. M. Buhmann. Towards weakly supervised semantic segmentation by means of multiple instance and multitask learning. In *CVPR*, 2010. 3, 5

[40] A. Vezhnevets, V. Ferrari, and J. M. Buhmann. Weakly supervised structured output learning for semantic segmentation. In *CVPR*, 2012. 3

[41] Y. Wei, X. Liang, Y. Chen, Z. Jie, Y. Xiao, Y. Zhao, and S. Yan. Learning to segment with image-level annotations. *Pattern Recognition*, 2016. 1, 3, 7

[42] T. Wu and S.-C. Zhu. A numerical study of the bottom-up and top-down inference processes in and-or graphs. *IJCV*, 93(2):226–252, 2011. 3

[43] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *ECCV*, 2014. 3

[44] J. Zhang, Z. Lin, J. Brandt, X. Shen, and S. Sclaroff. Top-down neural attention by excitation backprop. In *ECCV*, 2016. 2, 3, 4, 6

[45] W. Zhang, S. Zeng, D. Wang, and X. Xue. Weakly supervised semantic segmentation for social images. In *CVPR*, 2015. 1, 3

[46] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. Torr. Conditional random fields as recurrent neural networks. In *ICCV*, 2015. 1, 5

[47] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016. 3, 7