# Box2Seg: Attention Weighted Loss and Discriminative Feature Learning for Weakly Supervised Segmentation

Viveka Kulharia[⋆2], Siddhartha Chandra[*1],
Amit Agrawal[1], Philip Torr[2], Ambrish Tyagi[1]

[1]Amazon Lab126, [2]University of Oxford

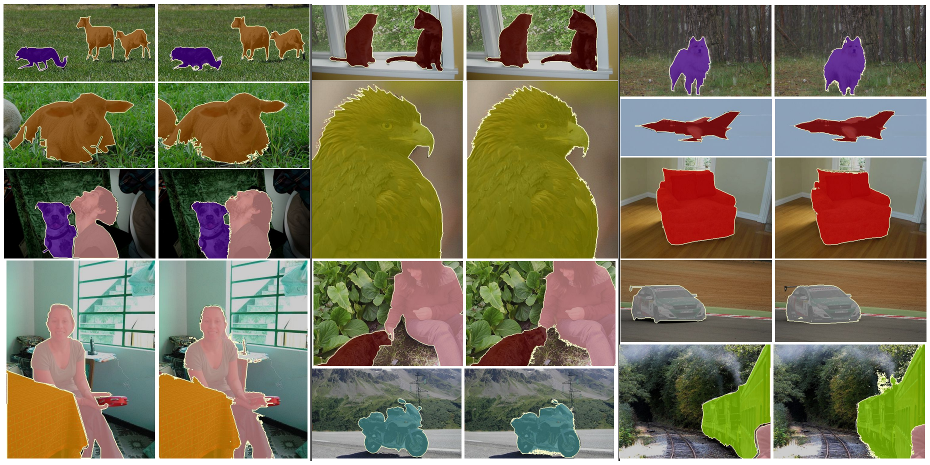

Fig. 1: Our Box2Seg model is able to produce high quality semantic segmentations using only bounding box annotations. Each image pair shows ground-truth (left) and predicted segmentation (right).

**Abstract.** We propose a weakly supervised approach to semantic segmentation using bounding box annotations. Bounding boxes are treated as noisy labels for the foreground objects. We predict a per-class attention map that saliently guides the per-pixel cross entropy loss to focus on foreground pixels and refines the segmentation boundaries. This avoids propagating erroneous gradients due to incorrect foreground labels on the background. Additionally, we learn pixel embeddings to simultaneously optimize for high intra-class feature affinity while increasing discrimination between features across different classes. Our method, Box2Seg, achieves state-of-the-art segmentation accuracy on PASCAL VOC 2012 by significantly improving the mIOU metric by 2.1% compared to previous weakly supervised approaches. Our weakly supervised approach is comparable to the recent fully supervised methods when fine-tuned with limited amount of pixel-level annotations. Qualitative results and ablation studies show the benefit of different loss terms on the overall performance.

[⋆] Authors contributed equally. V. Kulharia was an intern at Amazon Lab126.

# 1 Introduction

The accuracy of semantic segmentation approaches has improved significantly in recent years [8,10,32,35,53,54,56]. The mean Intersection-over-Union (mIoU) metric on the PASCAL VOC semantic segmentation benchmark has improved by over $20\%$ in the last five years. The success of these efforts can be broadly attributed to (i) advancements in deep neural network architectures and loss functions, (ii) efficient processing (better GPUs), and (iii) the availability of large datasets of images with human labeled per-pixel annotations [13,30]. Improvements in network architectures and hardware capabilities benefit all deep learning tasks. However, large datasets with per-pixel semantic labels are both expensive and slow to obtain [4,13] (typically 4-10 minutes per image), making it challenging to scale to a large number of object categories. Consequently, even the largest semantic segmentation datasets [30,58] include less than a couple of hundred object categories.

To address the scarcity of labeled data, some previous works have used synthetic datasets [43]. While labeling synthetically generated datasets involves little annotation effort, models trained on them do not always generalize well to the real world due to the domain gap between the real and the synthetic images. Alternative training strategies such as semi-, self-, or weak-supervision that require simpler/fewer labels (eg. image-level labels or bounding box annotations) have also been proposed. In this work, we show how to leverage real images with weak supervision to advance the state-of-the-art (SOTA) for semantic segmentation (refer to Fig. 1 for sample results).

Bounding box annotations yield high quality ground truth at a small cost. According to [4,13], per-pixel labeling takes over 4 minutes per image compared to $\sim 7$ seconds (35x faster) for annotating bounding boxes [36]. Furthermore, large datasets with bounding box annotations containing over 9 million images are publicly available [25]. Weakly supervised approaches using bounding box annotations have been shown to be more accurate when compared to methods that use only image-level labels. SOTA segmentation result on VOC using bounding box annotations [46] outperforms the SOTA methods using image-level labels [27] by $\sim 4\%$. Our work, *Box2Seg*, builds upon previous approaches that use bounding box annotations.

A key intuition of our paper is to consider bounding box annotations as containing *label noise* for the foreground object. Since bounding box annotation is a super-set of the actual object segmentation, this label noise is one-sided. In other words, some foreground labels are incorrectly assigned to background pixels within the bounding box. However, all foreground pixels inside the bounding box and background pixels outside all boxes are correctly labeled. Typical fully supervised segmentation training considers the label for every pixel as correct and gradients are back-propagated from all pixels. This would be an issue for a weakly supervised segmentation algorithm. To handle this, our algorithm predicts a novel per-pixel class-specific attention map and pixel embeddings in addition to the per-pixel segmentation output. The attention map is used to modulate the per-pixel cross entropy loss to handle label noise and reduces propagation of incorrect gradients. Thus, the attention map allows us to automatically discover salient regions of the object within the bounding box. The attention map is regularized using a soft filling-rate constraint for each training image.

We learn discriminative feature embeddings to capture long-range pairwise relationships between pixels across an image. We pretrain these embeddings to maximize the affinity between pixels belonging to same classes, while at the same time increasing the distance between features corresponding to different classes. During training, we define a novel loss function on pairs of pixels such that it encourages the pixel affinities to align with the predicted segmentation probabilities. A few methods have also proposed using feature affinity as an explicit measure to improve segmentation in fully-supervised and co-segmentation settings [5,18,22,31,33]. Affinities have also been used in some weakly supervised approaches [1,2], but they are trained in a fully-supervised manner using pseudo ground truth derived from class activation maps, as opposed to our embeddings which are trained by minimizing disagreements between them and estimated segmentation output (Sect. 3.4). We show how discriminative feature learning obtained using the model predictions can also be incorporated and is helpful in the context of weakly supervised semantic segmentation.

The remainder of this paper is organized as follows. We discuss related work in Sect. 2. Our learning algorithm and loss functions are described in Sect. 3. We demonstrate SOTA results on the PASCAL VOC 2012 segmentation benchmark (Sect. 4), outperforming previous weakly-supervised approaches by $2.1\%$ on the mIoU metric. We also show that our weakly supervised model serves as a good starting point for semi-supervised learning tasks, surpassing fully supervised baselines pre-trained with ImageNet [12] with only a fraction of pixel-level segmentation annotations.

## 2 Related Work

Previous works on weakly-supervised semantic segmentation have used image-level annotations [17,20,26,27,42,50,52], points/clicks [4], scribbles [29,47,48,49], bounding box annotations [11,19,37,41,46,55] and adversarial training [3,21]. We take a closer look at some of these methods and categorize them based on the labels required and their methodology.

**Image Level Labels:** Deep learning approaches that use image-level labels typically train a classification model first to recover coarse class activation maps [57], which describe class-discriminative image regions. The predicted class activation maps are then used as 'seeds' for optimization methods that grow the coarse activation maps to larger pseudo segmentation maps. A number of optimization methods such as superpixelization [26], deep seeded region growing [20], conditional random fields [42,52], and combinatorial grouping [40] have been proposed. Finally, the pseudo segmentation maps are used as ground truth to train the segmentation model [28]. Some approaches additionally employ a class-agnostic saliency estimation model [6,50,52] to capture *objectness* of pixels, and others employ Expectation-Maximization (EM) [17] to iteratively refine the pseudo ground truth and the parameters of the segmentation model.

**Bounding Box Labels:** The availability of bounding box annotations alleviates the need to estimate class activation maps for localizing objects of interest. The bounding boxes serve as crude segmentation masks which are refined using heuristic cues [23] and graph based optimization algorithms such as GrabCut [44] or mean-field inference [24] on a densely-connected conditional random field [37,41,46]. Previous works have used the

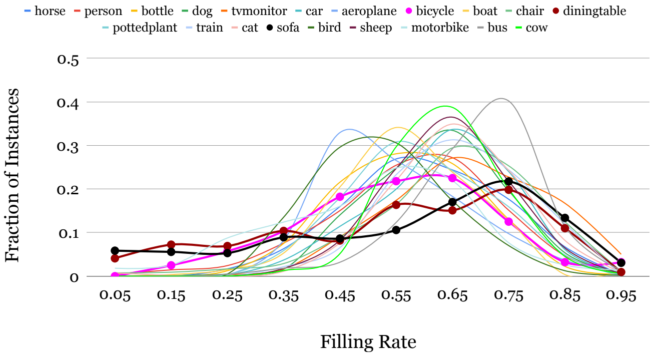

Fig. 2: Distribution of filling rate in PASCAL VOC 2012 training dataset. The per-class filling rates vary widely, especially for the categories such as sofa, dining table and bicycle. Unlike [46] that proposes using 1-3 filling rates per class, we use a per-image class specific filling rate.

refined masks for training the segmentation model [28]. Some of these approaches additionally use EM [11,23,37,55] for iterative refinement of the ground truth and model parameters. However, training models with EM is time-consuming and is prone to propagation of errors over iterations; In contrast, our proposed approach does not require iterative refinement and outperforms these methods by a large margin (Sect. 4).

**Multi-tasking**: Hu *et al.* [19] train an object detector and segmentation model simultaneously using the Mask-RCNN framework [15] assuming a closed form relationship between parameters of the detection and segmentation branches. While [19] is able to benefit from the advantages of multi-tasking, it requires (a) (limited) training data with per-pixel annotations, (b) a region of interest (RoI) proposal method, and (c) uses the RoI warping module which captures local context alone. Our approach does not require any per-pixel annotations and uses a fully-convolutional network. We capture global image context via long-range interactions by training pixel embeddings.

**Label Noise**: We consider bounding boxes as noisy labels for foreground objects. Some fully-supervised approaches [39,34] have tackled noise in the segmentation ground truth by consolidating predictions from two or more independent classifiers and discarding pixels with ambiguous predictions from the training data. Among weakly supervised methods, Song *et al.* [46] propose the idea of using the filling-rate as a cue to supervise training with bounding box ground truth. The filling-rate of a class [46] is defined as the average proportion of pixels in a bounding box that instances of the class occupy, and is estimated by applying dense-CRF [24] for foreground extraction within bounding boxes. For example, suppose in the PASCAL VOC 2012 dataset, instances of 'sheep' occupy roughly $60\%$ of pixels in each bounding-box. During training, their approach ignores gradients from the $40\%$ pixels with the lowest confidences in each bounding box containing 'sheep'. In contrast, we advocate estimating a *per-image class-specific* spatial attention map, to allow for large intra-class variations in object appearance (see Fig. 2). Instead of ignoring loss on pixels with low-confidence as in [46], which is a hard constraint, we use the filling rate to regularize the attention map as a soft constraint. Therefore, our attention map offers a continuous modulation of cross-entropy loss, rather than a binary decision. Attention maps have also been employed in [1,59] where they are trained in a fully-supervised manner using pseudo segmentation ground truth. In contrast, our attention modulated loss offers a principled way of handling label noise present in segmentation ground truth derived from bounding box annotations. We validate our approach with quantitative and qualitative comparisons with [46] in Sect. 4.

## 3 Proposed Approach

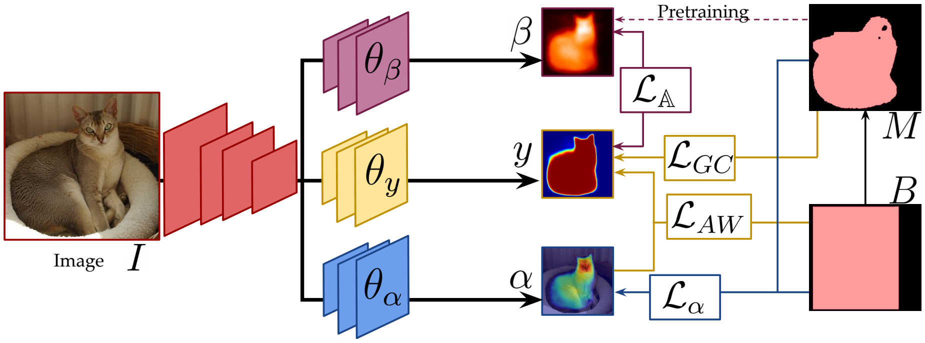

Fig. 3: Overview of our weakly supervised Box2Seg training pipeline. Our segmentation model is a three branch CNN which outputs segmentation probabilities $y$, per-class spatial attention maps $\alpha$, and pixel embeddings $\beta$. $B$ denotes segmentation masks defined using bounding box annotations. GrabCut is applied to $B$ to obtain refined segmentation masks $M$. We use $B$ and $M$ to supervise the three outputs of the model. Pixel embeddings capture long-range pairwise relationships and the attention map refines the segmentation output $y$ by reducing the effect of label noise. At inference time we only use the branch producing the segmentation probabilities $y$, and discard the other two branches that output $\alpha, \beta$.

In this section, we describe our Box2Seg algorithm in detail. We discuss our pipeline and loss functions that allow us to learn a per-image, class-specific attention map as well as pixel embeddings.

### 3.1 Feed Forward Network Architecture

Let $I \in \mathbb{R}^{n \times 3}$ denote a $3-$channel input color image with $n$ pixels. Our segmentation model $\mathcal{S}$ is a fully-Convolutional Neural Network (CNN) which takes $I$ as input and produces three outputs: (i) the segmentation output probabilities $y \in \mathbb{P}^{m \times (L+1)}, \mathbb{P} \in [0, 1]$, (ii) the attention map $\alpha \in \mathbb{P}^{m \times L}$, and (iii) the pixel embeddings $\beta \in \mathbb{R}^{m \times d}$. Here $d$ denotes the size of the pixel embeddings and $m$ denotes the spatial resolution of the outputs. Each pixel in the output image can assume one of the $L + 1$ labels ($L$ object categories and the background class). The ground truth bounding boxes are denoted by $B_{box} \in \mathbb{R}^{K \times 5}$, where $K$ is the number of bounding boxes comprising of 4 coordinates and a class label. To simplify the notation, we denote by $B \in \mathbb{R}^{m\times(L+1)}$ the box-segmentation tensor obtained by setting all pixels inside a bounding box to 1 for the corresponding class label (channel). In the case two boxes overlap, we assign 1 to the class corresponding to the smaller bounding box (assuming that the smaller box is in the front). Note that this assumption may not always be true and can result in incorrect label assignments for $B$ (See eg. Fig. 8).

Similar to prior works [37,41,46], we generate pseudo ground-truth segmentation maps $M \in \mathbb{R}^{m\times(L+1)}$ by applying classical graph-based unsupervised segmentation approach (Grabcut [44]) on each bounding box in our training dataset. The segmentation masks obtained by classical methods, albeit noisy and imprecise, provide a good prior for training deep learning models.

Fig. 3 gives an overview of our approach. Training involves passing the input image $I$ through a common feature encoder that feeds into the three branches of network $\mathcal{S}$ to produce $y$, $\alpha$, and $\beta$, as follows:

$$y = \mathrm{Softmax}(\mathcal{S}(I,\theta_y));\ \ \alpha = \sigma(\mathcal{S}(I,\theta_\alpha));\ \ \beta = \mathcal{S}(I,\theta_\beta),$$

where $\theta_y, \theta_\alpha, \theta_\beta$ denote the parameters of the model $\mathcal{S}$ for the respective branches. $\mathrm{Softmax}$ denotes the softmax over all the classes and $\sigma$ denotes the sigmoid activation function. Note that the segmentation output probabilities, $y$, sums up to 1 for each pixel across classes due to softmax. However, the activation maps, $\alpha$, use sigmoid output, making them independent for each class. Note that our model does not use additional parameters compared to any baselines we compare against in Sec. 4 as we discard the branches producing $\alpha, \beta$ at test time.

### 3.2 Box And GrabCut Based Losses

We use the box-segmentation tensor $B$ to train a simple baseline by minimizing the following cross-entropy loss:

$$\mathcal{L}_{box} = -\frac{1}{m}\sum_{c=0}^{L}\sum_{i=1}^{m} B(i,c)\log\left(y(i,c)\right). \tag{1}$$

Similar to previous works, we use the GrabCut outputs $M$ obtained from the bounding boxes to define another baseline by minimizing the following cross-entropy loss:

$$\mathcal{L}_{GC} = -\frac{1}{m}\sum_{c=0}^{L}\sum_{i=1}^{m} M(i,c)\log\left(y(i,c)\right). \tag{2}$$

Since the GrabCut algorithm provides reasonable segmentation outputs (Sect. 4), we use $\mathcal{L}_{GC}$ in addition to our loss functions described in the following sections.

### 3.3 Attention Weighted Segmentation Loss

Our novel attention modulated cross-entropy loss considers bounding box annotations as *noisy labels* for the foreground object. Note that since the bounding box is a superset of the actual object segmentation mask, the label noise is one-sided: foreground labels are incorrectly assigned to background pixels, but no true foreground labels are missing. Additionally, pixels outside all the bounding boxes can be considered *definite background* and do not have any label noise. Since supervised segmentation training typically considers all labels as correct, in the presence of label noise, erroneous gradients can be back propagated during training. At pixels close to bounding box and object boundaries, the network gets conflicting information about the foreground/background labels at similar pixels. This is the reason for the worse performance of baseline trained with box annotations only (Table 3).

**Attention on Foreground Objects:** We propose to modulate the per-pixel cross-entropy loss using the predicted attention map from the network by minimizing

$$\mathcal{L}_{fg} = \frac{-1}{\sum_i^m B(i,c)} \sum_{c=1}^{L} \sum_{i=1}^{m} \alpha(i,c) B(i,c) \log\left(y(i,c)\right). \tag{3}$$

Note that the attention map has same spatial resolution as the segmentation output and is class-specific. The attention weighted loss is only defined for the $L$ foreground classes. In addition, the loss is normalized with the size of the bounding box, to give similar weighting to each class.

**Background Loss:** Since the pixels outside all the bounding boxes can be considered as definite background, the background loss is defined as

$$\mathcal{L}_{bg} = -\frac{\sum_i^m B(i,0)\log(y(i,0))}{\sum_i^m B(i,0)}, \tag{4}$$

where $0$ denotes the background class. We define the **A**ttention **W**eighted **L**oss (AWL) as

$$\mathcal{L}_{AW} = \mathcal{L}_{fg} + \mathcal{L}_{bg}. \tag{5}$$

**Attention Map Regularization:** Without any regularization on the attention maps, the network can minimize $\mathcal{L}_{fg}$ by predicting all $\alpha(i,c) = 0$. To prevent this, we compared two approaches to regularize the attention map. The first approach regularizes the attention mask using ground truth bounding boxes (similar to [46]), by minimizing the $\mathcal{L}_2$ loss between the attention maps and the bounding boxes.

$$\mathcal{L}_{\alpha}^{bbox} = \sum_{i=1}^{m} \sum_{c=1}^{L} \parallel B(i,c) - \alpha(i,c) \parallel^2 . \tag{6}$$

The second approach regularizes the attention mask using fill-ratios obtained from GrabCut outputs $M$ as follows. Let $\eta_c$ denote the per class, per image, filling rate defined as the proportion of pixels in the pseudo ground-truth $M$ compared to its corresponding bounding box.

$$\eta_c = \frac{\sum_i^m M(i,c)}{\sum_i^m B(i,c)}. \tag{7}$$

Similarly, the predicted fill rate of the attention map is computed as

$$\eta_c^{'} = \frac{\sum_i^m B(i,c)\alpha(i,c)}{\sum_i^m B(i,c)}. \tag{8}$$

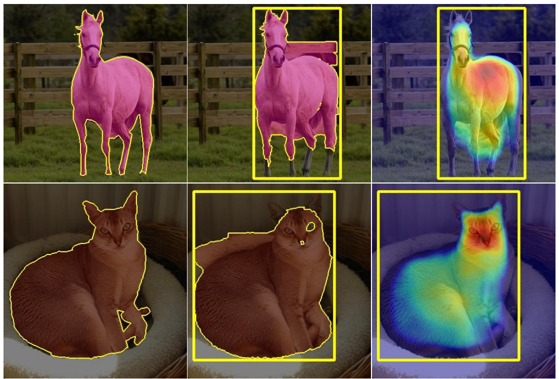
(a) Image + GT (b) Grabcut + Box (c) Attention + Box

Fig. 4: Visualizing attention maps produced by our network. (a) Input images with overlaid ground truth masks. (b) Grabcut output with ground-truth bounding box. (c) Predicted attention map. Note that while Grabcut output is erroneous and often includes background pixels, our attention maps are concentrated on the objects of interest.

We use a margin loss to ensure that the predicted fill rate is at least a factor $\gamma$ of the fill rate obtained using $M$.

$$\mathcal{L}_{\alpha}^{fr} = \max(0, \gamma\eta_c - \eta_c^{'}), \tag{9}$$

where $\gamma \in [0, 1]$ is a hyper-parameter which is set to $0.7$ in our experiments. Thus, the predicted fill rate (Eqn. 8) is allowed to vary between $\gamma\eta_c$ and 1. Equation 9 enforces a soft constraint that the attention map should allow propagation of loss from at least $70\%$ of $\eta_c$ pixels inside the bounding box. Using $L_{\alpha}^{bbox}$ for regularization forces the attention mask to take the shape of the bounding box. Thus, it is prone to include background pixels in attention map. Using $L_{\alpha}^{fr}$ provides a softer constraint and gives better results in our experiments. Figure 4 shows some qualitative examples where attention map is able to focus on foreground pixels despite errors in the underlying GrabCut segmentations, allowing Box2Seg to be robust to label noise.

### 3.4 Discriminative Feature Learning

We now describe supervision of our pixel embeddings $\beta$ which capture long-range pairwise relationships between different pixels. Our pixel embeddings can be denoted as $\beta = \{\beta_i\}$, where $\beta_i$ is the $d$-dimensional feature for the $i^{th}$ pixel. Affinity between embeddings at pixel $i$ and $j$ is given by its normalized dot product

$$\mathbb{A}(i, j) = \beta_i \cdot \beta_j = \frac{\beta_j^T \beta_i}{\|\beta_j\| \|\beta_i\|}. \tag{10}$$

Intuitively, we want to achieve high affinity between feature vectors of two pixels that belong to the same class, while ensuring low affinity between features of two different classes. Similarly, a background pixel should have low affinity with respect to

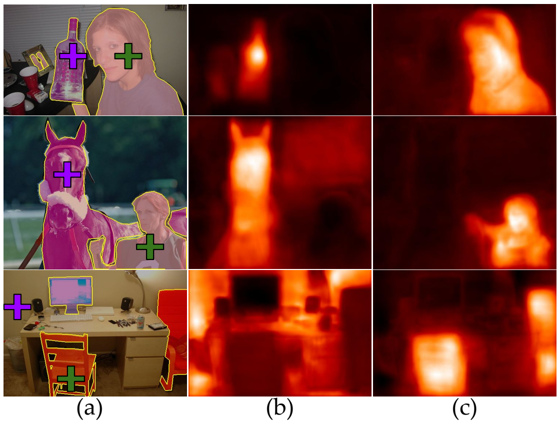

Fig. 5: Visualization of affinities produced by our network. (a) Input image with overlaid ground truth masks. Two pixels are marked with **purple** and **green** + signs. (b,c) Heatmap of affinities (Eqn. 10) w.r.t. to the **purple** and **green** pixels, respectively.

another pixel that belongs to one of the $L$ foreground classes. To achieve this, we define a novel loss function on pairs of pixels ($i$,$j$), such that it encourages the pixel affinities to align with the predicted segmentation probabilities, as follows,

$$\mathcal{L}_{\mathbb{A}} = \sum_{i,j} \left(\mathbb{A}(i,j) - y_j^T y_i\right)^2 . \tag{11}$$

However, training affinity matrices requires large amount of memory. To avoid creating large affinity matrices of size $m \times m$, we randomly sample a small fraction of pixel empeddings equally from each class to compute this loss. Figure 5 shows the affinity maps computed from our class discriminative embeddings.

### 3.5 Training Box2Seg

Our approach optimizes the following loss function:

$$\mathcal{L} = \mathcal{L}_{\mathcal{GC}} + \lambda_{AW}\mathcal{L}_{AW} + \lambda_{\alpha}\mathcal{L}_{\alpha} + \lambda_{\mathbb{A}}\mathcal{L}_{\mathbb{A}}, \tag{12}$$

where $\mathcal{L}_{\alpha}$ equals either $\mathcal{L}_{\alpha}^{bbox}$ or $\mathcal{L}_{\alpha}^{fr}$. $\lambda_{AW}$, $\lambda_{\alpha}$ and $\lambda_{\mathbb{A}}$ are weights applied to the individual losses.

## 4 Experiments and Results

We evaluate the performance of our approach on PASCAL VOC 2012 [13] dataset. Our ablation studies provide insights into our design choices. Finally, we demonstrate that our method provides a better initialization than Imagenet pretraining for the task of semi-supervised segmentation.

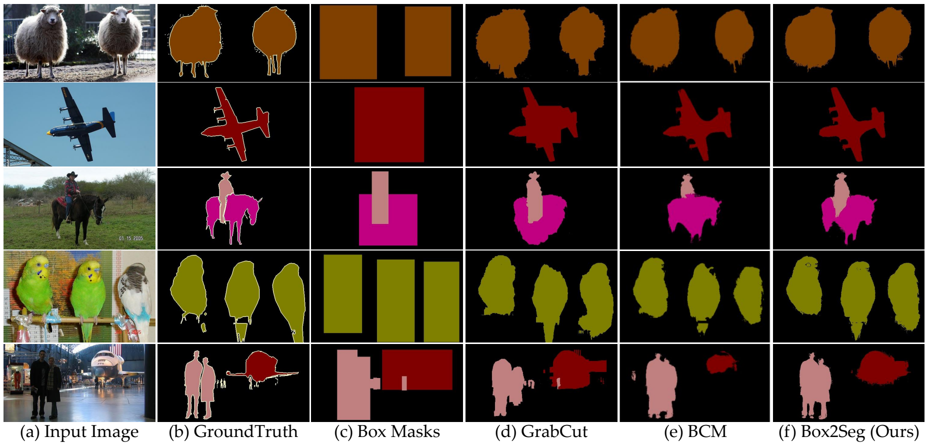

Fig. 6: Comparison of our segmentation results with those of BCM [46].

### 4.1 Implementation Details

Our segmentation network architecture is similar to UPerNet [51] where the encoder backbone is ResNet-101 [16], and decoders consist of 2 convolutional layers. We employ the ResNet-101 backbone to ensure fair comparison with the two most recent SOTA works, SDI [23] and BCM [46] as well as 4 other recent methods [27,47,48,49] in Table 1. We have three decoders, one each for the $y$, $\alpha$, and $\beta$ branches. The final results are spatially down-sampled by a factor of 4, i.e. $m = n/16$. We start with ImageNet pretrained [12] weights to initialize our encoder.

We train our network in two stages. First, we pre-train the feature representations using the affinity loss by randomly sampling 10% of pixel pairs using the Grabcut outputs M. This is done by minimizing Eqn. 11 using $M_i$ in place of $y_i$. Note that the pretraining phase is meant to only serve the task of weight initialization for our discriminative feature learning. Our final feature representations are robust to noise in the Grabcut outputs since they are trained eventually to agree with our predictions using Eqn. 11. After the pretraining phase, we enable the decoder branch to also output $y$ and $\alpha$ and train the entire network end-to-end to optimize the loss function in Eqn. 12. After doing a grid search of hyper-parameters on a held out validation set, $\lambda_{AW}$, $\lambda_\alpha$ and $\lambda_{\mathbb{A}}$ are set to 10, 1, and 1 respectively. We use Stochastic Gradient Descent to train our models for 40 epochs with an initial learning rate $= 1e-4$, momentum $= 0.9$, weight decay $= 5e^{-4}$ and a polynomially decaying learning rate as in [7]. Our implementation uses PyTorch [38] and is trained on Nvidia's TitanX GPUs. Note that at test time we discard the decoders yielding $\alpha, \beta$, therefore our method uses no additional parameters compared to any of our baselines in the following sections. Our implementation will be available at www.github.com/vivkul/Box2Seg.

### 4.2 Quantitative and Qualitative Evaluation

PASCAL VOC 2012 is one of the gold standard benchmarks for semantic segmentation. Following [9,11,32], we use the augmented annotation set [14] consisting of 10582

| **Method** | **Annotations** | **Backbone** | **mIoU** |
|---|---|---|---|
| GrabCut-NoTrain-GT | box | - | 71.6 |
| GrabCut-NoTrain-Det | box | - | 68.5 |
| SSNet [52] | image-level | DenseNet-169 | 63.3 |
| F2FA [27] | image-level | ResNet-101 | 66.5 |
| ScribbleSup (C) [29] | scribble | VGG-16 | 63.1 |
| NormalCut [47] | scribble | ResNet-101 | 72.8 |
| BPG [49] | scribble | ResNet-101 | 73.2 |
| KernelCut [48] | scribble | ResNet-101 | 73.0 |
| WSSL (C) [37] | box | VGG-16 | 60.6 |
| BoxSup (C) [11] | box | VGG-16 | 62.0 |
| SDI (C) [23] | box | ResNet-101 | 69.4 |
| BCM (C) [46] | box | ResNet-101 | 70.2 |
| Li et al. (C) [28] | box | ResNet-101 | 74.3 |
| **Box2Seg** | box | ResNet-101 | **74.9** |
| **Box2Seg** (C) | box | ResNet-101 | **76.4** |

Table 1: Comparison of Box2Seg to previous weakly supervised semantic segmentation methods on PASCAL VOC validation set. C=dense-CRF post processing.

training and 1449 validation images. Performance is measured using the mIoU metric on the validation set.

We compare the accuracy of our algorithm to the SOTA weakly supervised segmentation methods on VOC validation set in Table 1. Our approach shows a large improvement over previous methods ([37,11,23,46,28]), some of which have also used VGG-16 backbones for feature representation. Li *et al.* [28] report an mIoU of 74.3%, using bounding boxes as supervision and dense CRF post-processing. Our method yields an mIoU of 74.9% without dense CRF, and 76.4% mIoU after dense CRF post processing resulting in an improvement of 2.1%. Box2Seg also outperforms recent weakly supervised methods that use image level labels [27,52] or scribbles [29,47,48,49]. Per-category results comparing Box2Seg against the SOTA methods are reported in Table 2.

Qualitative comparison of our results with BCM [46] is shown in Fig. 6. Our method is able to produce higher-quality object segmentations compared to BCM. It's also robust to false segmentations in some cases, e.g. false detection closer to the left edge of

| Method | bkg | aero | bike | bird | boat | bottle | bus | car | cat | chair | cow | table | dog | horse | mbike | person | plant | sheep | sofa | train | tv | mIoU |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| BCM VGG (CRF) [46] | 89.8 | 68.3 | 27.1 | 73.7 | 56.4 | 72.6 | 84.2 | 75.6 | 79.9 | 35.2 | 78.3 | 53.2 | 77.6 | 66.4 | 68.1 | 73.1 | 56.8 | 80.1 | 45.1 | 74.7 | 54.6 | 66.8 |
| BCM ResNet (CRF) [46] | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | 70.2 |
| Li et al. (CRF) [28] | 93.3 | 85.0 | 35.9 | 88.6 | 70.3 | 77.9 | 91.9 | 83.6 | 90.5 | 39.2 | 84.5 | 59.4 | 86.5 | 82.4 | 81.5 | 84.3 | 57.0 | 85.9 | 55.8 | 85.8 | 70.4 | 75.7 |
| Box2Seg | 92.5 | 66.5 | 31.7 | 78.9 | 65.5 | 83.4 | 90.4 | 86.7 | 86.0 | 55.1 | 81.8 | 59.9 | 80.5 | 74.1 | 76.0 | 75.7 | 65.3 | 85.1 | 72.5 | 87.8 | 77.7 | 74.9 |
| Box2Seg (CRF) | 93.3 | 72.4 | 33.0 | 84.2 | 64.9 | 83.5 | 90.9 | 86.7 | 88.7 | 57.2 | 83.6 | 62.5 | 82.6 | 76.8 | 77.0 | 77.8 | 63.3 | 87.2 | 75.1 | 88.3 | 74.1 | **76.4** |

Table 2: Per-class results on PASCAL VOC 2012 Validation set. We compare our *Box2Seg* results (with and without denseCRF) with those of the previous state-of-the-art methods. Please note that Li et al. (CRF) results require COCO annotations. Per-class results from [46] for their ResNet-101 model are not available.

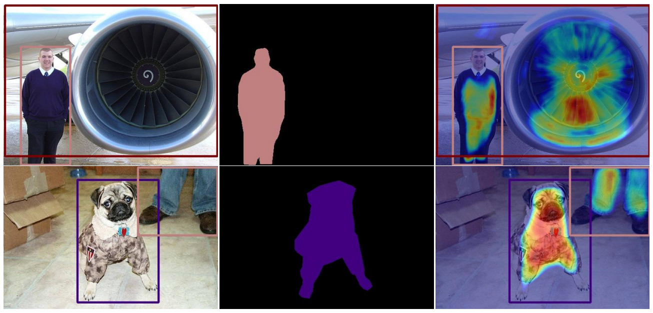

Fig. 7: (Left) Input images with overlaid bounding box labels. (Middle) Incorrect pixel-level annotations on foreground objects such as *Aeroplane* and *Person*. Note that the bounding box annotations are correct. (Right) Predicted attention map with bounding boxes. Please refer to Sect. 4.4 for discussion on how our AWL can help in such cases with conflicting ground truth during fully-supervised training.

the BCM result on the last row are suppressed by our method. Please refer to Fig. 1 for additional examples of segmentations using our approach. We also show some failure cases of our approach in Fig. 8.

**Accuracy of GrabCut (no training):** We also evaluated the accuracy of GrabCut algorithm itself, without any training, against the segmentation ground-truth. Interestingly, GrabCut output on *ground truth* bounding boxes (*GrabCut-NoTrain-GT*) results in a strong weakly-supervised baseline in itself with $71.6\%$ mIoU. However, since the ground truth bounding boxes are not available at inference, a more practical baseline is to obtain bounding boxes on the validation set using an object detector (we used SNIPER [45]) and then run GrabCut on those. This baseline is referred to as *GrabCut-NoTrain-Det* and obtains $68.5\%$ mIoU.

### 4.3 Ablation studies

Table 3 shows the efficacy of our loss functions in improving the performance of our approach. The trivial *Box* baseline obtained by training the segmentation network with bounding box supervision ($\mathcal{L}_{box}$ loss only) results in a low mIoU of 59.3%, as expected.

| **Method** | $\mathcal{L}_{box}$ | $\mathcal{L}_{GC}$ | $\mathcal{L}_{AW}$ | $\mathcal{L}_{\alpha}$ | $\mathcal{L}_{\mathbb{A}}$ | **mIoU** |
|---|---|---|---|---|---|---|
| Box | ✓ | | | | | 59.3 |
| GrabCut | | ✓ | | | | 72.7 |
| Affinity | | ✓ | | | ✓ | 73.9 |
| AW-box | | ✓ | ✓ | $\mathcal{L}_{\alpha}^{bbox}$ | | 74.1 |
| AW-fr | | ✓ | ✓ | $\mathcal{L}_{\alpha}^{fr}$ | | 74.6 |
| Box2Seg | | ✓ | ✓ | $\mathcal{L}_{\alpha}^{fr}$ | ✓ | **74.9** |

Table 3: Ablation study showing the effect of our loss functions in improving the performance over baseline methods.

| Method | CE Loss | + $\mathcal{L}_{AW}$ | $\boldsymbol{\Delta}$ |
|---|---|---|---|
| Supervised baseline | 73.6 | 75.1 | +1.5 |

Table 4: Improvements in segmentation accuracy on the PASCAL VOC 2012 validation set in the fully supervised setting using AWL.

Training the segmentation network with *GrabCut* masks as supervision ($\mathcal{L}_{GC}$ loss only) without affinity or attention losses resulted in 72.7% mIOU. Introducing our feature embeddings to the pipeline (Affinity) improves the mIoU to 73.9%. Our novel AWL terms ($\lambda_{AW}\mathcal{L}_{AW} + \lambda_{\alpha}\mathcal{L}_{\alpha}$) significantly improve the mIoU to 74.6% with filling-rate regularization (*AW-fr*) (more about AWL in Sect. 4.4). Finally, combining all losses (Eqn 12), *Box2Seg* obtains 74.9% mIoU.

### 4.4 Effectiveness of Attention Weighted Loss in the Fully-Supervised Setting

In this section, we demonstrate that AWL can be used in the fully-supervised setting, and it boosts segmentation accuracy when we have disagreements between the bounding-box and per-pixel annotations.

During our analysis, we found that roughly 10% of training images in PASCAL VOC dataset have disagreements between the bounding box annotations and pixel-level annotations. Fig. 7 shows few examples, where pixel-level annotation for object categories such as *Aeroplane* and *Person* are missing, but their corresponding bounding box labels are correctly provided. Due to incorrect pixel-level annotations, fully supervised training would back-propagate erroneous gradients on these pixels. Since our novel AWL ($\mathcal{L}_{AW}$ in Sect. 3.3) is effective in dealing with label noise for weakly supervised networks, we analyze if it can further improve the performance of a fully supervised network also. In cases where conflicting sources of ground truth exist (as in Fig. 7), AWL can allow correct gradients to propagate back due to correct bounding box annotations (see the predicted attention maps in Fig. 7), thereby reducing the effect of incorrect pixel-level annotations.

Table 4 shows the improvement obtained by adding AWL to the fully *Supervised baseline*. Adding the AWL (+ $\mathcal{L}_{AW}$) to the fully-*Supervised* baseline improved the segmentation accuracy by 1.5% in mIoU. Thus, AWL is effective at improving segmentation accuracy, both in the weakly- and fully- supervised settings.

### 4.5 Semi-supervised Semantic Segmentation

Weakly supervised trained method can naturally serve as a starting point for semi-supervised segmentation. To study this, we fine tuned our Box2Seg model using different amount of pixel-level segmentation annotations. We observe significant improvements in accuracy, even with small amount of supervision as shown in Table 5. For comparison, we show the result of semi-supervised fine tuned BCM [46] model trained with 1464 images (13.8% of the data) and another fully supervised baseline (*DeepLab*) using ImageNet as the initialization. Our weakly supervised baseline results improved from 74.9% mIoU to 83.1% with just 10% of supervised data. Therefore, our Box2Seg model can serve as a good starting point and provides better results compared to ImageNet based initialization.

| Method | **mIoU** |
|---|---|
| BCM [46] † (S=0%) (C) | 70.2 |
| BCM [46] † (S=13.8%) (C) | 71.6 |
| DeepLab † (S=100%) (C) | 74.5 |
| **Box2Seg** (S=0%) | 74.9 |
| **Box2Seg** (S=5%) | 78.7 |
| **Box2Seg** (S=10%) | 83.1 |
| **Box2Seg** (S=100%) | 86.4 |

Table 5: Semi-supervised segmentation using Box2Seg model as initialization. (S=$\tau$%) implies $\tau$% images using pixel annotations are used for semi-supervised fine-tuning. C=dense-CRF post processing. †: results using ResNet-101 backbone taken from [46].

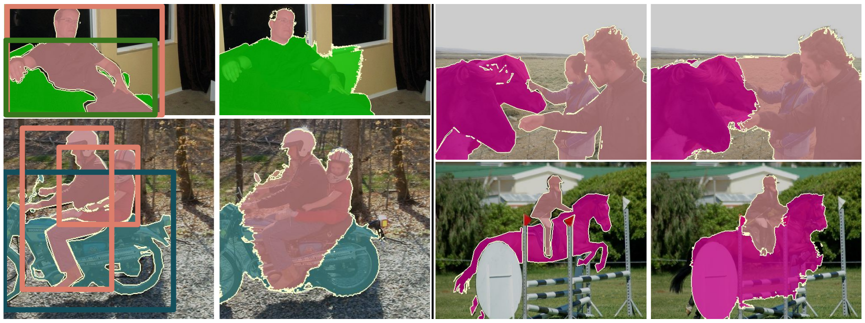

Fig. 8: Failure Cases. (Left) Incorrect label assignment on box annotations $B$ can happen when the smaller bounding box is physically behind the larger bounding box. (Right) Predicted segmentation bleeds into background.

## 5 Conclusions

In this work, we proposed a pipeline for training a weakly supervised semantic segmentation method from bounding box annotations. We showed that bounding box annotations can be treated as noisy labels for foreground objects and proposed a novel attention weighted loss to reduce the effect of erroneuos gradients due to incorrect labels. We also proposed pixel embeddings to capture global context via long-range pairwise interactions. We showed qualitative improvements over the previous SOTA on the PASCAL VOC semantic segmentation benchmark and pushed the mIoU metric forward by 2.1%. Interestingly, fully supervised methods can also benefit from attention weighted loss in the presence of exhaustive bounding box annotations but missing pixel-level annotations. Future work would involve using an edge detector as a cue to learn class boundaries, and also extending our method to benefit from image-level supervision.

***Acknowledgements:*** V. Kulharia worked as an intern at Amazon Lab126 and continued this work at Oxford University. V Kulharia's DPhil is funded by the Toyota Research Institute grant.

## References

1. Ahn, J., Cho, S., Kwak, S.: Weakly supervised learning of instance segmentation with inter-pixel relations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2209–2218 (2019) 3, 4
2. Ahn, J., Kwak, S.: Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018) 3
3. Arandjelović, R., Zisserman, A.: Object discovery with a copy-pasting gan. arXiv preprint arXiv:1905.11369 (2019) 3
4. Bearman, A., Russakovsky, O., Ferrari, V., Fei-Fei, L.: What's the point: Semantic segmentation with point supervision. In: European conference on computer vision (ECCV). pp. 549–565. Springer (2016) 2, 3
5. Chandra, S., Usunier, N., Kokkinos, I.: Dense and low-rank gaussian crfs using deep embeddings. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5103–5112 (2017) 3
6. Chaudhry, A., Dokania, P.K., Torr, P.H.: Discovering class-specific pixels for weakly-supervised semantic segmentation. In: British Machine Vision Conference (BMVC) (2017) 3
7. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Semantic image segmentation with deep convolutional nets and fully connected crfs. arXiv preprint arXiv:1412.7062 (2014) 10
8. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. arXiv:1606.00915 (2016) 2
9. Chen, L., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. CoRR **abs/1706.05587** (2017), http://arxiv.org/abs/1706.05587 10
10. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: European Conference on Computer Vision (ECCV). pp. 801–818 (2018) 2
11. Dai, J., He, K., Sun, J.: Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In: International Conference on Computer Vision (ICCV). pp. 1635–1643 (2015) 3, 4, 10, 11
12. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2009) 3, 10
13. Everingham, M., Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International Journal on Computer Vision (IJCV) **88**(2), 303–338 (Jun 2010). https://doi.org/10.1007/s11263-009-0275-4, http://dx.doi.org/10.1007/s11263-009-0275-4 2, 9
14. Hariharan, B., Arbeláez, P., Girshick, R., Malik, J.: Hypercolumns for object segmentation and fine-grained localization. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015) 10
15. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. International Conference on Computer Vision (ICCV) (2017) 4
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016) 10
17. Hou, Q., Massiceti, D., Dokania, P.K., Wei, Y., Cheng, M.M., Torr, P.H.: Bottom-up top-down cues for weakly-supervised semantic segmentation. In: International Workshop on

Energy Minimization Methods in Computer Vision and Pattern Recognition. pp. 263–277. Springer (2017) 3

18. Hsu, K.J., Lin, Y.Y., Chuang, Y.Y.: Deepco3: Deep instance co-segmentation by co-peak search and co-saliency detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 3
19. Hu, R., Dollár, P., He, K., Darrell, T., Girshick, R.: Learning to segment every thing. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4233–4241 (2018) 3, 4
20. Huang, Z., Wang, X., Wang, J., Liu, W., Wang, J.: Weakly-supervised semantic segmentation network with deep seeded region growing. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 7014–7023 (2018) 3
21. Hung, W.C., Tsai, Y.H., Liou, Y.T., Lin, Y.Y., Yang, M.H.: Adversarial learning for semi-supervised semantic segmentation. In: British Machine Vision Conference (BMVC) (2018) 3
22. Ke, T.W., Hwang, J.J., Liu, Z., Yu, S.X.: Adaptive affinity fields for semantic segmentation. In: European Conference on Computer Vision (ECCV) (2018) 3
23. Khoreva, A., Benenson, R., Hosang, J., Hein, M., Schiele, B.: Simple does it: Weakly supervised instance and semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 876–885 (2017) 3, 4, 10, 11
24. Krähenbühl, P., Koltun, V.: Efficient inference in fully connected crfs with gaussian edge potentials. In: Neural Information Processing Systems (NIPS) (2011) 3, 4
25. Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Duerig, T., et al.: The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. arXiv preprint arXiv:1811.00982 (2018) 2
26. Kwak, S., Hong, S., Han, B.: Weakly supervised semantic segmentation using superpixel pooling network. In: AAAI Conference on Artificial Intelligence (2017) 3
27. Lee, J., Kim, E., Lee, S., Lee, J., Yoon, S.: Frame-to-frame aggregation of active regions in web videos for weakly supervised semantic segmentation. In: International Conference on Computer Vision (ICCV) (October 2019) 2, 3, 10, 11
28. Li, Q., Arnab, A., Torr, P.H.: Weakly-and semi-supervised panoptic segmentation. In: Proc. European Conference on Computer Vision (ECCV 18). pp. 102–118 (2018) 3, 4, 11
29. Lin, D., Dai, J., Jia, J., He, K., Sun, J.: Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3159–3167 (2016) 3, 11
30. Lin, T., Maire, M., Belongie, S.J., Bourdev, L.D., Girshick, R.B., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: common objects in context. CoRR **abs/1405.0312** (2014), http://arxiv.org/abs/1405.0312 2
31. Liu, S., De Mello, S., Gu, J., Zhong, G., Yang, M.H., Kautz, J.: Learning affinity via spatial propagation networks. In: Neural Information Processing Systems (NIPS) (2017) 3
32. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3431–3440 (2015) 2, 10
33. Maire, M., Narihira, T., Yu, S.X.: Affinity cnn: Learning pixel-centric pairwise relations for figure/ground embedding. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016) 3
34. Min, S., Chen, X., Zha, Z.J., Wu, F., Zhang, Y.: A two-stream mutual attention network for semi-supervised biomedical segmentation with noisy labels (2018) 4
35. Noh, H., Hong, S., , Han, B.: Learning deconvolution network for semantic segmentation. In: arXiv preprint arXiv:1505.04366 (2015) 2

36. Papadopoulos, D.P., Uijlings, J.R., Keller, F., Ferrari, V.: Extreme clicking for efficient object annotation. In: IEEE International Conference on Computer Vision (CVPR). pp. 4930–4939 (2017) 2
37. Papandreou, G., Chen, L.C., Murphy, K.P., Yuille, A.L.: Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1742–1750 (2015) 3, 4, 6, 11
38. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch. Pytorch (2017) 10
39. Petit, O., Thome, N., Charnoz, A., Hostettler, A., Soler, L.: Handling missing annotations for semantic segmentation with deep convnets. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, pp. 20–28. Springer (2018) 4
40. Pont-Tuset, J., Arbelaez, P., Barron, J.T., Marques, F., Malik, J.: Multiscale combinatorial grouping for image segmentation and object proposal generation. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **39**(1), 128–140 (2016) 3
41. Rajchl, M., Lee, M.C., Oktay, O., Kamnitsas, K., Passerat-Palmbach, J., Bai, W., Damodaram, M., Rutherford, M.A., Hajnal, J.V., Kainz, B., et al.: Deepcut: Object segmentation from bounding box annotations using convolutional neural networks. IEEE transactions on medical imaging **36**(2), 674–683 (2016) 3, 6
42. Redondo-Cabrera, C., Baptista-Ríos, M., López-Sastre, R.J.: Learning to exploit the prior network knowledge for weakly supervised semantic segmentation. IEEE Transactions on Image Processing **28**(7), 3649–3661 (2019) 3
43. Richter, S.R., Hayder, Z., Koltun, V.: Playing for benchmarks. International Conference on Computer Vision (ICCV) (2017) 2
44. Rother, C., Kolmogorov, V., Blake, A.: Grabcut: Interactive foreground extraction using iterated graph cuts. In: ACM transactions on graphics (TOG). vol. 23, pp. 309–314. ACM (2004) 3, 6
45. Singh, B., Najibi, M., Davis, L.S.: SNIPER: Efficient multi-scale training. Neural Information Processing Systems (NIPS) (2018) 12
46. Song, C., Huang, Y., Ouyang, W., Wang, L.: Box-driven class-wise region masking and filling rate guided loss for weakly supervised semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3136–3145 (2019) 2, 3, 4, 5, 6, 7, 10, 11, 13, 14
47. Tang, M., Djelouah, A., Perazzi, F., Boykov, Y., Schroers, C.: Normalized cut loss for weakly-supervised cnn segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1818–1827 (2018) 3, 10, 11
48. Tang, M., Perazzi, F., Djelouah, A., Ben Ayed, I., Schroers, C., Boykov, Y.: On regularized losses for weakly-supervised cnn segmentation. In: European Conference on Computer Vision (ECCV). pp. 507–522 (2018) 3, 10, 11
49. Wang, B., Qi, G., Tang, S., Zhang, T., Wei, Y., Li, L., Zhang, Y.: Boundary perception guidance: a scribble-supervised semantic segmentation approach. In: International Joint Conference on Artificial Intelligence (IJCAI) (2019) 3, 10, 11
50. Wang, X., You, S., Li, X., Ma, H.: Weakly-supervised semantic segmentation by iteratively mining common object features. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1354–1362 (2018) 3
51. Xiao, T., Liu, Y., Zhou, B., Jiang, Y., Sun, J.: Unified perceptual parsing for scene understanding. In: European Conference on Computer Vision (ECCV). pp. 418–434 (2018) 10
52. Zeng, Y., Zhuge, Y., Lu, H., Zhang, L.: Joint learning of saliency detection and weakly supervised semantic segmentation. International Conference on Computer Vision (ICCV) (2019) 3, 11

53. Zhang, H., Dana, K., Shi, J., Zhang, Z., Wang, X., Tyagi, A., Agrawal, A.: Context encoding for semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018) 2
54. Zhang, H., Zhang, H., Wang, C., Xie, J.: Co-occurrent features in semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 2
55. Zhao, X., Liang, S., Wei, Y.: Pseudo mask augmented object detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4061–4070 (2018) 3, 4
56. Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D., Huang, C., Torr, P.: Conditional random fields as recurrent neural networks. In: International Conference on Computer Vision (ICCV) (2015) 2
57. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2921–2929 (2016) 3
58. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017) 2
59. Zhu, Y., Zhou, Y., Xu, H., Ye, Q., Doermann, D., Jiao, J.: Learning instance activation maps for weakly supervised instance segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3116–3125 (2019) 4