Pseudo-Mask Augmented Object Detection

Xiangyun Zhao
UCSD

Jifeng Dai
Microsoft Research

Nuno Vasconcelos
UCSD

Yichen Wei
Microsoft Research

Abstract

In this work, we present a novel and effective framework to facilitate object detection with the instance-level segmentation information that is only supervised by bounding box annotation. Starting from the joint object detection and instance segmentation network, we propose to recursively estimate the pseudo ground-truth object masks from the instance-level object segmentation network training, and then enhance the detection network with top-down segmentation feedbacks. The pseudo ground truth mask and network parameters are optimized alternatively to mutually benefit for each other. To obtain the promising pseudo masks in each iteration, we embed a graphical inference that incorporates the low-level image appearance consistency and the bounding box annotations to refine the segmentation masks predicted by the segmentation network. Our approach progressively improves the object detection performance by incorporating the detailed pixel-wise information learned from the weakly-supervised segmentation network. Extensive evaluation on the detection task in PASCAL VOC 2007 and 2012 [11] verifies that the proposed approach is effective and obtains state-of-the-art performance.

1. Introduction

Recent years have seen significant progresses in object detection. Since the deep convolutional neutral network has been firstly used in R-CNN [18], a lot of improvements have been made, and they improve the performance from many different aspects, *e.g.*, deeper networks and stronger features [38, 39, 22], better object proposals [32, 40], more discriminative and powerful features [25, 1], more accurate localization [30, 14].

In this work, we investigate the object detection task from another important aspect, that is, how to exploit object segmentation to improve object detection. Although it has been well recognized in the literature that the two tasks are closely related and detection could benefit from segmentation, most previous works, *e.g.*, [1, 7], share two common drawbacks.

First, they rely on accurate and pixel-wise ground truth segmentation masks for the segmentation problem. However, such mask annotation is very expensive to obtain. Instead, most large-scale object recognition datasets such as ImageNet [9] and PASCAL VOC [11] only provide bounding box level annotations. In addition, most of these methods only explore how to facilitate object detection with semantic image segmentation, which did not independently consider the characteristics of each instance. We argue that the instance-level segmentation task is more aligned with object detection by considering the object information from different granularity (pixel-level versus box-level). How to effectively use only bounding boxes for the segmentation to facilitate detection remains an unclear problem.

Second, most works have independent network structures for segmentation and detection tasks, *e.g.*, the state-of-the-art MNC [7]. Although the two tasks often share the same underlying convolutional features, the two networks do not directly interact with each other and the commonality between the two tasks may not be fully exploited. For the existing approaches, the benefits of jointing learning are mostly from the better learned deep feature representation as in a normal multi-task setting. It is seldom explored that how segmentation information can benefit detection directly and more closely in a deep learning framework.

In this work, we propose a novel approach that better addresses the above two issues, which augments the object detector with generated object masks from the bounding box annotation, named as Pesudo-mask Augmented Detection (PAD). It starts from a strong baseline network architecture that directly integrates the state-of-the-art Fast-RCNN [17] network for object detection and InstanceFCN [5] for object segmentation, in a normal multi-task setting.

Given the baseline network, we make two major contributions. First, our PAD treats ground truth object segmentation masks as hidden variables as they are unknown, which are gradually refined by only using bounding box annotations as the supervision, called as *pseudo ground truth* *masks*. The pseudo masks of training images and the network parameters are optimized alternatively in an EM-like way. To make the alternative learning more effective, we propose two novel techniques. Between each iteration, the pseudo masks are progressively refined by embedding a graphical model, which improves the pixel-wise estimation with a graph-cut optimization with low-level appearance coherence and the ground truth bounding boxes as additional constraints. Beside the iteratively refined pseudo masks, we also incorporate a novel 1D box loss defined over the groundtruth box, as a supervision signal to help improve quality of pseudo masks learning, similar to LocNet [16].

Second, based on the commonality of segmentation and detection tasks, as well as the correlations of the network structures, we propose to connect the two networks such that the segmentation information provides a top-down feedback for detection network. In this way, the learning of detection network is improved as additional supervision signals are back propagated from the segmentation branch. The top-down segmentation feedback considers two contexts, on both the *the global* level and *instance* level. Their effectivenesses on improving detection accuracy are both verified in experiments.

The proposed approach is validated using various state-of-the-art network architectures (VGG and ResNet) on several well-known object detection benchmarks (i.e., PASCAL VOC 2007 and 2012). The strong and state-of-the-art performance verifies its effectiveness.

2. Related Work

Joint Segmentation and Detection There exists quite a few works that integrate the object segmentation and detection tasks [13, 19, 10, 4, 1, 7]. In spite of their various techniques, these methods have common limitations: 1) the pixel-level segmentation annotation is required, which is difficult to obtain, 2) the integration of segmentation and detection is usually loose due to the separately trained segmentation and detection network. Our work overcomes the two limitations in an integrated learning framework, where the top-down segmentation feedback is proposed to bridge the segmentation and detection network.

Using Graphical Models for Segmentation Graphical models are widely used for traditional image and object segmentation [3, 33, 27, 26, 35, 24]. Compared to the feature representation learning by the CNNs, the graphical inferences possess the merits of effectively incorporating local and global image constraints (*e.g.* appearance consistencies, and structure priors) into a single optimization framework. Recently, some recent works integrate graphical models (e.g., CRF/MRF) into the deep neutral networks [41, 29] for a joint training.

In our approach, traditional graph cut based optimization [3] is embedded to refine the pseudo ground truth mask estimation during the iterative learning. It effectively refines the quality of pixel-wise pseudo masks to progressively improve the discriminative capability of detection and segmentation network.

Weakly Supervised Segmentation Due to the difficulty of obtaining large-scale pixel-wise segmentation ground truth, some works resort to weakly supervised learning of segmentation, such as using bounding box annotation [6] or scribbles [28]. Such methods share some similarity with ours by using the iterative optimization to gradually refine the segmentation. However, different from their focus on single image segmentation, our approach jointly optimizes the detection and weakly-supervised object segmentation network, with more complex and effective learning technique.

3. Pseudo-mask Augmented Detection

We focus on facilitating the object detection with the instance-level object segmentation information, using only ground truth bounding box annotations. We denote the set of all ground truth boxes as $B^{gt} = \{B_o^{gt}\}$, where subscript o enumerates all objects in all training images. We use the former notation throughout the paper for its simplicity.

As motivated earlier, it is beneficial to estimate the per-pixel object segmentation as well. An auxiliary object segmentation task is added in a normal multi-task setting. That is, the two tasks share the same underlying convolutional feature maps. Since the ground truth binary object segmentation masks are unknown, we treat them as hidden variables, which are first initialized with B^{gt}, and then iteratively refined in our approach. We call them estimated object masks as *pseudo ground truth masks* from the bounding box annotation, denoted as $M^{pseudo} = \{M_o^{pseudo}\}$.

Let the network parameters be Θ, and the network output for object segmentation and detection be $M(\Theta)$ and $B(\Theta)$, respectively. The network parameters are learned to minimize the loss function

$$L_{seg}(M(\Theta)|M^{pseudo}, B^{gt}) + L_{det}(B(\Theta)|B^{gt}), \qquad (1)$$

where the two loss terms are enforced on object segmentation and detection tasks, respectively. As defined the network optimization target, the performance of detection network heavily depends on the quality of estimated pseudo masks M^{pseudo}. That is, the poor estimation of M^{pseudo} leads to poor network learning of $M(\theta)$, which in turn would cause negative chain effect on whole iterative framework for object detection. We propose an effective learning approach that progressively improves the quality M^{pseudo}

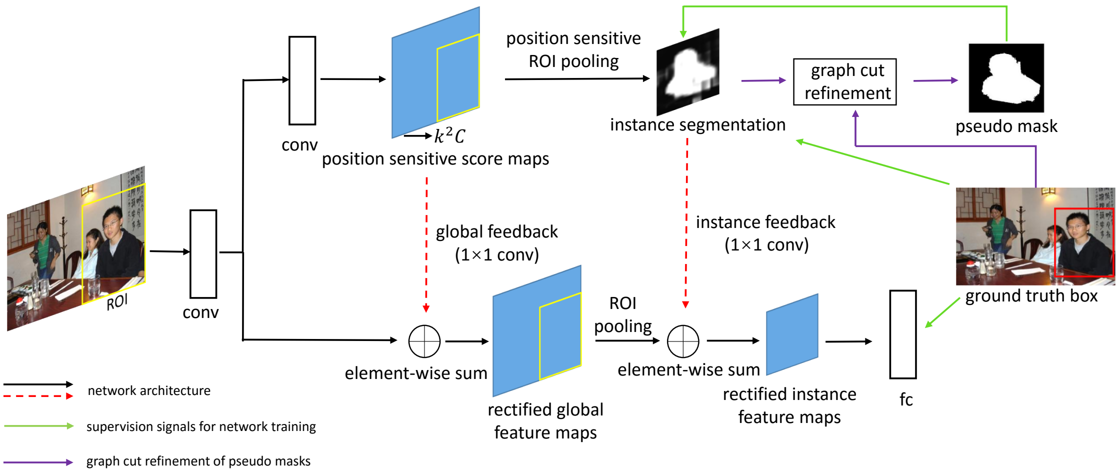
position sensitive ROI pooling
instance segmentation
graph cut refinement
pseudo mask
conv
k^2C
position sensitive score maps
ROI
conv
global feedback (1×1 conv)
instance feedback (1×1 conv)
ground truth box
element-wise sum
ROI pooling
element-wise sum
rectified global feature maps
rectified instance feature maps
fc
network architecture
supervision signals for network training
graph cut refinement of pseudo masks

Figure 1. An overview of our pseudo-mask augmented object detection, consisting of the network architecture and graph cut based pseudo mask refinement. For each image, the detection sub-network and instance-level object segmentation sub-network share convolutional layers (i.e. conv1-conv5 for VGG, conv1-conv4 for ResNet). For segmentation sub-network, position-sensitive score maps are generated by 1×1 convolutional layer, and it is then passed through position-sensitive pooling to obtain object masks. The predicted object masks and bounding box annotations are then combined to refine pseudo masks with a graph-cut refinement, which also provide the supervision signal for network training in next iteration. In each iteration, we explore the global feedback and instance top-down feedback from the instance segmentation sub-network to facilitate the object detection sub-network for better detection performance.

from a coarse initialization using B^{gt}, as summarized in Algorithm 1. The detection network parameters Θ and pseudo masks M^{pseudo} are alternatively optimized following a EM-like way, with the other fixed in each iteration.

Note that Algorithm 1 only operates on training images. The learned network parameters Θ are applied on test images to generate detection and segmentation results.

The instance-level segmentation masks $M(\Theta)$ from the pixel-wise prediction of segmentation network are usually noisy and poor. This is partially because pseudo masks are not accurate enough, and the estimation is made in a pixel-wise manner, which does not consider the correlations between the pixels such as smoothness constraints used in most segmentation approaches. As shown in Algorithm 1, we thus propose two novel ingredients to achieve the effective iterative learning. First, in each object mask refinement step (Sec. 3.2), the pseudo ground truth mask for each object is improved using the traditional graphical inference. It is formulated as a global optimization problem that considers not only the current mask estimation from the network, but also the low level image appearance coherence and the ground truth bounding boxes, which is efficiently solved by graph cut [2].

Second, we notice that only using the pseudo mask M^{pseudo} as 2D pixel-wise supervision signals may be not sufficient as the masks themselves are often noisy and not accurate enough. Thus, the 1D box loss(explained in Sec. 3.1) in Eq. (1) and (2) (Sec. 3.1) is incorporated to consider the additional constraints provided by the ground truth bounding box. The 1D loss term complements the noisy 2D segmentation loss and performs better regularization on the segmentation network learning.

Algorithm 1 Iterative learning of network parameters Θ and pseudo ground truth masks M^{pseudo}.

```
**input**: ground truth bounding boxes $B^{gt}$
initialize the pseudo masks $M^{pseudo}$ from $B^{gt}$;
learn $\Theta_0$ with loss in Eq. (1)                        ▷ Sec. 3.1
**for** $t = 1$ **to** $T$ **do**
    refine $M^{pseudo}$ from $M(\Theta_{t-1})$ and $B^{gt}$   ▷ Sec. 3.2
    learn $\Theta_t$ with loss in Eq. (1)                    ▷ Sec. 3.1
**end for**
**output**: final network parameters $\Theta_t$
**output**: pseudo ground truth masks $M^{pseudo}$
```

With the aforementioned two novel techniques, both pseudo masks M^{pseudo} and network parameters Θ are improved steadily, benefiting from each other. Based on the refined object masks, we add connections between the segmentation and detection sub-networks such that the segmentation features provide top-down feed back for the detection, leading to better results in the object detection (Sec. 3.1).

3.1. Network Architecture and Training

As shown in Figure 1, following the common multi-task learning, we adopt two sub-networks for the object segmentation and detection tasks, which are built on the shared feature maps. We first extract the object proposals, or regions of interest (ROIs) from the Region Proposal Network (RPN) [32]. For simplicity, we do not use the complex training strategy in [32]. Instead, we pre-train the RPN and fix

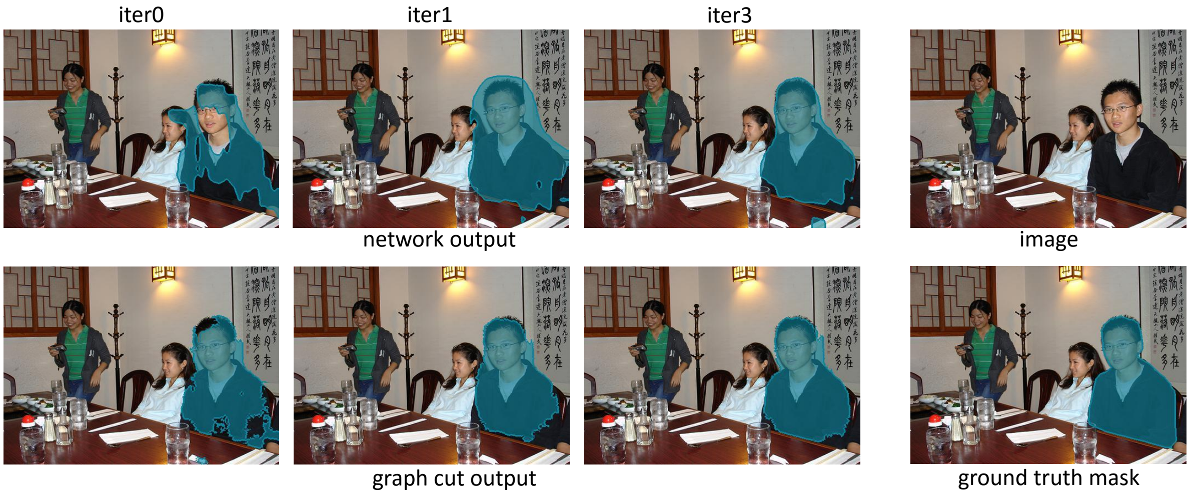

Figure 2. Example predicted pseudo masks by the iterative refinement and employing graph cut optimization.

the ROIs throughout our experiments.

Object Segmentation with Pseudo Masks In general, this sub-network can adopt any instance-level object segmentation network such as DeepMask [31], MNC [7], etc. In this work, we adopt the recent InstanceFCN [5], since it achieves state-of-the-art results on producing object segment proposals. It applies 1×1 convolutional layer on the feature maps to produce k^2 *positive sensitive* score maps. The k^2 (we use $k = 7$ in this work) maps encode the relative-position information between image pixels and ROIs (*e.g.*, top-left, bottom-right). Given a ROI, its per-pixel score map is assembled by dividing the positive sensitive score maps into k^2 cells and copying each cell's content from the corresponding k^2 maps. This step generates a fixed-size (28×28) per-ROI foreground probability map.

However, this strategy still has several limitations: first, it only runs on square sliding windows; second, it ignores object category information and is limited to object generate object segment proposals. We extend this approach in two aspects to seamlessly integrate it into the object detection network. First, we extend its k^2 score maps to $k^2 * C$ score maps, where C indicates the category number. In this way, the individual segmentation module for each category is optimized. Second, we employ generic object proposals [32] to replace the square sliding windows and the position-sensitive ROI pooling layer in [8] on the proposals. During training, a sigmoid layer is applied on each category of the per-ROI score maps to generate the instance foreground probability maps.

Segmentation Loss Our approach estimate a corresponding pseudo mask for each object instance. For a ROI that has intersection-over-union (IOU) larger than 0.5 with a ground-truth object, we define a per-pixel sigmoid cross-entropy loss on the ROI's foreground probability map, with respect to current pseudo mask of the object that is regarded as a hidden ground truth mask. We call this term *2D mask loss*.

Since the pseudo masks are quite noisy, it may damage our network if directly using it as the supervision. Since each ground truth bounding box tightly encloses the object, this implies that for a horizontal or vertical scan line in the box, at least one pixel on the line should be foreground. On the other hand, all pixels outside of the box should be background. Accordingly, we define a *1D box loss* term for each ROI. Specifically, the predicted foreground mask of each ROI is projected to two 1-d vectors along the horizontal and vertical directions, respectively, by applying a max operation on all values of a scan line. For each position on the 1-d vector, it is denoted as foreground when its corresponding line is inside the box. Otherwise, it is denoted as background. The 1D loss term summarizes the sigmoid cross-entropy loss on all positions of the 1-d vectors. Note that a similar 1D loss idea has also been utilized in LocNet [16] for bounding box regression. In this work, we use it for object segmentation.

In summary, by combining the 2D mask loss and 1D box loss, the segmentation loss in Eq. (1) is computed by

$$L_{\text{seg}} = L_{2D}(M(\theta)|M^{pseudo}) + L_{1D}(M(\theta)|B^{gt}). \quad (2)$$

Object Detection with Top-down Segmentation Feedback For the detection sub-network, we use the state-of-the-art Fast(er) R-CNN [17, 32]. It applies a ROI pooling layer for each ROI on the feature maps to obtain per-ROI feature maps, and then applies fully connected (FC) layers to output detection results.

In the common multi-task setting, the two sub-networks do not interact and only use the separate optimization target. However, the object segmentation and detection tasks are highly correlated to each other and the sub-networks also share similar structures, we connect the two so that the segmentation network provides top-down feedback information for detection network.

The feedback from segmentation consists of the *global feedback* and the *instance feedback* (the two red dotted arrows in Figure 1). In terms of the global level, the $k^2 * C$ position-sensitive score maps in the segmentation sub-network (before ROI pooling) encode the segmentation information on the whole image. They are of the same spatial dimension of the shared convolutional feature maps but different feature channels, in general. A 1×1 convolutional layer is applied on the score maps to change its channel number to match that of the shared feature maps. The two sets of maps are then element-wisely summed to produce the "rectified" global feature maps for the object detection sub-network.

In terms of instance level, the instance segmentation masks (i.e., per-ROI score maps) encode the specific pixel-wise characteristics of each object instances. As shown in Figure 1, after the ROI pooling step in both sub-networks, the per-ROI instance segmentation score maps from the segmentation branch are passed through a 1×1 convolutional layer and max pooling layer to obtain feature maps with the same dimension as the per-ROI feature maps of the detection branch. The score maps from two branches are then summed to produce the "rectified" instance feature maps.

Afterwards, several fully connected (FC) layers are used to generate the object classification scores and bounding box regression results, in the same way as Fast RCNN [17]. The detection loss in Eq. (1) includes the classification softmax loss and bounding box regression loss for all ROIs,

$$L_{det} = L_{cls}(B(\theta)|B^{gt}) + L_{reg}(B(\theta)|B^{gt}). \qquad (3)$$

Training Given the estimated pseudo masks M_{pseudo} in each step, the instance-level segmentation network and object detection network are optimized by stochastic gradient descent, using image centric sampling [17]. In each mini-batch, two training images are randomly sampled. The loss gradients from Eq. (1) are back propagated to update all the network parameters jointly.

3.2. Pseudo Mask Refinement

Accurate pseudo mask is the key to bridge the object detection and instance-level segmentation networks. The estimated pseudo masks directly from the segmentation network are usually noisy and blurred. More importantly, as the pixels are considered individually by the convolutional network, the informative interactions between pixels are not fully exploited, such as the smoothness constraint used in traditional image and object segmentation.

In this work, we explore a graphical model to refine the pseudo mask estimation, which jointly incorporates the current mask probabilities from the instance-level segmentation network, the low level image appearance cues and the ground-truth bounding box information. The graphical model is defined on a graph constructed by the super-pixels generated by [12] for each object instance. For each graph, a vertex denotes a super-pixel while an edge is defined over neighboring super-pixels. Note that in this step the spatial range of the pseudo mask is enlarged by 20% from the ground truth bounding box, in order to include more boundary areas and thus improve segmentation quality.

Formally, for all super-pixels $\{x_i\}$ in the pseudo mask under consideration, we estimate their binary labels $\{y_i\}$, where $y_i = 1$ indicates foreground, and 0 for background. Similar to the traditional object segmentation approaches [3, 33], we define a global objective function in the form of

$$\sum_i U(y_i) + \sum_{i,j} V(y_i, y_j). \qquad (4)$$

Unary Term. The unary term $U(y_i)$ measures the likelihood of the super pixel x_i being foreground. It considers both the foreground probabilities from the network and the ground truth bounding box b^{gt}, defined as

$$U(y_i) = \begin{cases} 0 & \text{if } y_i = 0 \text{ and } x_i \notin b^{gt} \\ \infty & \text{if } y_i = 1 \text{ and } x_i \notin b^{gt} \\ 1 - prob_{fg}(x_i) & \text{if } y_i = 0 \text{ and } x_i \in b^{gt} \\ prob_{fg}(x_i) & \text{if } y_i = 1 \text{ and } x_i \in b^{gt}. \end{cases} \qquad (5)$$

The first two cases ensure that x_i is background when it is outside the ground truth bounding box. The last two cases directly adopt the results from the current network estimation when x_i is inside, where $prob_{fg}(x_i)$ sums the foreground probabilities of all pixels within the super pixel x_i. To obtain a pixel's foreground probability, firstly we only consider the probability of the ground truth object category as the foreground probability. Secondly, the segmentation sub-network outputs all ROIs' mask probability maps. To obtain the foreground probability on a single ground truth object, we find all ROIs with IoU larger than 0.5 with the ground-truth object, and average their foreground probabilities together as the foreground probabilities for the object.

Pairwise Term. The pair-wise binary term $V(y_i, y_j)$ considers the local smoothness constraints, defined on all neighboring super pixels. It uses the low level image cues similarly as in [3, 33]. If the neighboring super-pixels are similar in appearance, the cost of assigning them different labels should be high. Otherwise, the cost is low.

We use both color and texture information to measure the similarity as in [28]. For a super-pixel x_i, its color histogram $h_c(x_i)$ is built on the RGB color space using 25 bins

method	w/ seg. loss	w/ mask refinement	w/ global feedback	w/ instance feedback	VGG-16	ResNet-50	ResNet-101
ION [1]					75.6		
CRAFT [40]					75.7		
HyperNet [25]					76.3		
Faster-RCNN [32]					73.2	74.9	76.4
Ours (a)					74.5	78.1	79.4
Ours (b)	✓				74.9	78.2	79.8
Ours (c)	✓	✓			75.7	78.6	79.9
Ours (d)	✓	✓	✓		76.0	78.7	80.0
Ours (e)	✓	✓		✓	76.2	78.9	80.6
Ours (f)	✓	✓	✓	✓	**77.0**	**79.6**	**80.9**

Table 1. Object detection results on VOC 2007 test training on the union set of VOC 2007, VOC 2012 train and validation dataset

for each channel. The texture histogram $h_t(x_i)$ is built on the gradients at the horizontal and vertical orientations with 10 bins for each. The pair-wise binary term is defined as

$$\begin{aligned} V(y_i, y_j) =& [y_i \neq y_j] \Big\{ - \frac{\|h_c(x_i) - h_c(x_j)\|_2^2}{\delta_c^2} \\ & - \frac{\|h_t(x_i) - h_t(x_j)\|_2^2}{\delta_t^2} \Big\}, \end{aligned} \tag{6}$$

where $[\cdot]$ is 1 or 0 if the subsequent argument is true/false. δ_c and δ_t are set as 5 and 10, respectively.

The objective function in Eq. (4) is minimized by graph cut solver [2], for the pseudo mask of each object instance in each image. The resulting binary labels $\{y_i\}$ define the refined binary pseudo masks M_{pseudo}, which are then used as supervision signals to train the network in next iteration (Algorithm 1).

Some exemplar predicted pseudo masks of object instances are illustrated in Figure 2. It can be observed that the graph cut can help generate more complete pseudo masks and preserve better boundary information. Moreover, the pseudo masks can also be effectively refined with more iterative steps, benefiting from both the segmentation and detection network.

4. Experiments

4.1. Implementation and Training Details

Our experiments are based on Caffe [23] and public Faster-RCNN code [32]. For simplicity, the region proposal network (RPN) is trained once and the obtained object proposals are fixed. We evaluate the performance of the proposed PAD using three state-of-the-art network structures: VGG-16 [37], ResNet-50 [22] and ResNet-101 [22]. We use the publicly available pre-trained model on ILSVRC2012 [34] to initialize all network parameters. Following the practice in [17, 22, 8], the ROI pooling layer is inserted between the convolutional and the fc layers for VGG-16, and between conv4 and conv5 blocks of convolutional layers for ResNet models.

Each mini-batch contains 2 randomly selected images, and we sample 64 region proposals per image leading to 128 ROIs for each network updating step. To reduce the effect of randomization on the detection performance, we fix the random seed so that all training architectures take the training images in the same order. After training the baseline Faster R-CNN model [32] with OHEM [36] using the above settings, we actually obtain better accuracy than that reported in the origianl Faster R-CNN [32], as also revealed in Table 1 .

We run SGD for 80k iterations with learning rate 0.001 and 40k iterations with learning rate 0.0001. The iteration number T in Algorithm 1 is set as 3 since no further performance increase is observed. Detection accuracies in the following sections are measured by the standard mean average precision (mAP) scores.

4.2. Ablation study on VOC 2007

Table 1 compares different strategies and variants in our proposed approach, as well as the results from representative state-of-the-art works as reference. Following the protocol in [17], all models are trained on the union set of VOC 2007 [11] *trainval* and VOC 2012 *trainval*, and are evaluated on VOC 2007 *test* set. We evaluate results using VGG-16 [38], ResNet-50 [22], and ResNet-101 models.

We start from the baseline where no pseudo mask is used (Table 1(a)). This is equivalent to our faster R-CNN implementation, which sets a strong and clean baseline. It achieves 74.5%, 78.1%, and 79.4% mAP scores by using VGG-16, ResNet-50, and ResNet-101 models, respectively.

We then evaluate the naive pseudo mask baseline (Table 1(b)). It simply uses pseudo masks initialized from the bounding boxes, without iterative refinement. The joint training is performed on both the object detection loss and the segmentation loss w.r.t. the naive pseudo masks. This is equivalent to a simple multi-task baseline with coarse pseudo masks. It obtains slightly higher accuracies than (a). It indicates that multi-task learning is slightly helpful but limited, as pseudo mask quality is very poor.

Method	**Train**	**mAP**
HyperNet [25]	07++12	71.4
CRAFT [40]	07++12	71.3
MR-CNN [15]	07++12	73.9
Faster-RCNN(VGG16) [32]	07++12	70.4
Faster-RCNN(ResNet100)	07++12	73.8
PAD (VGG-16)	07+12	**74.4**
PAD (ResNet-101)	07+12	**79.5**

Table 2. Detection results on VOC 2012 test. 07+12: 07 trainval + 12 trainval, 07++12: 07 trainvaltest + 12 trainval.

To evaluate iterative mask refinement, we report the results of our variant that does not use the mask feedback for object detection network, as Table 1(c). It differs from Table 1(b) in that the iterative mask refinement is used. After three iterations, the mAP scores are 75.7%, 78.6%, and 79.9%, respectively, which are 1.2%, 0.5%, and 0.5% higher than Table 1(a). It verifies that our approach is capable of generating more reasonable pseudo masks.

Furthermore, after performing the top-down segmentation feedback, the detection performance can be further improved by comparing the results of Table 1(f) and Table 1(c). Our full model achieves mAP scores of 77.0%, 79.6%, and 80.9% using different networks, respectively, which are 2.5%, 1.5%, and 1.5% higher than Table 1(a). The strong results validate the advantages of using top-down segmentation feedback to explicitly model the interaction between segmentation and detection networks.

To disentangle the effectiveness of the global and instance feedbacks, we block one of them respectively in Table 1(d) and Table 1(e). We observe that both the feedbacks are effective for boosting the object detection accuracy, and combining them achieves the largest gain.

4.3. Detection Results on VOC 2012

Our approach is also compared with the state-of-the-art models on comp4 (outside data) track on PASCAL VOC 2012 evaluation server. The training data is the union of VOC 2007, VOC 2012 train and validation dataset, following [17]. As reported in Table 2[1], our approach obtains 74.4% and 79.5% with VGG-16 and ResNet-101, which are substantially better than currently leading methods.

4.4. Segmentation and Detection on VOC 2012 SDS

To validate the quality of the estimated pseudo masks, we evaluate the performance of Simultaneous Detection and Segmentation (SDS) task on VOC 2012 SDS set [21], where the ground-truth instance segmentation masks are available.

method	train w/ gt mask?	mAP^r (%)	mAP^b (%)
Faster R-CNN		-	65.5
MNC [7]	✓	63.5	-
PAD w/ gt mask	✓	**64.5**	**68.1**
PAD w/ Grabcut mask		48.3	66.9
PAD w/o 1D box loss		49.1	66.9
PAD iter. 0		44.3	66.7
PAD iter. 1		52.1	67.0
PAD iter. 2		58.0	67.5
PAD		58.5	67.6

Table 3. Performance comparison on VOC 2012 SDS task.

This subset is widely used to evaluate instance-aware segmentation methods [21, 7]. Following the protocols in [21, 7], the model training and evaluation are performed on 5623 images from VOC 2012 *train*, and 5732 images from VOC 2012 *validation* sets, respectively. The ground-truth instance-level segmentation masks are provided by the additional annotations from [20]. The mask-level mAP^r scores and the box-level mAP^b scores are employed as the evaluation metrics for instance-level mask estimation and object detection performance measure, respectively.

First, we report the upper-bound result of training the PAD model using ground-truth object masks, i.e., PAD w/ gt mask in Table 3). It achieves 68.1% mAP^b and 64.5% mAP^r. We note that this "oracle" upper bound is strong. The segmentation accuracy is even higher than the state-of-the-art instance-aware segmentation method, MNC [7].

Second, we evaluate the superiority of alternative training for pseudo mask estimation and network. As shown in Table 3, PAD obtains mAP^b 67.6% and mAP^r 58.5%, which are slightly worse than the upper bound. The instance-level object segmentation results steadily improve with more iterative refinements (iter.1, iter.2). The iterative improvement can be observed in Figure 2. Finally, PAD w/ Grabcut mask shows the instance-level object segmentation results by using the pseudo masks obtained by Grabcut method [33], which is a traditional state-of-the-art object segmentation method. PAD w/o 1D box loss corresponds to results without using the 1D box loss as in Eq. 2. Their comparison with PAD in Table 3 demonstrate the effectiveness of iterative graph cut refinement and 1D segmentation loss, the key indigents of PAD.

Our approach obtains good results for simultaneous object detection and segmentation. Examples are show in Figure 3 and 4.

[1] http://host.robots.ox.ac.uk:8080/anonymous/PTJSWL.html, http://host.robots.ox.ac.uk:8080/anonymous/KZ-LBX.html

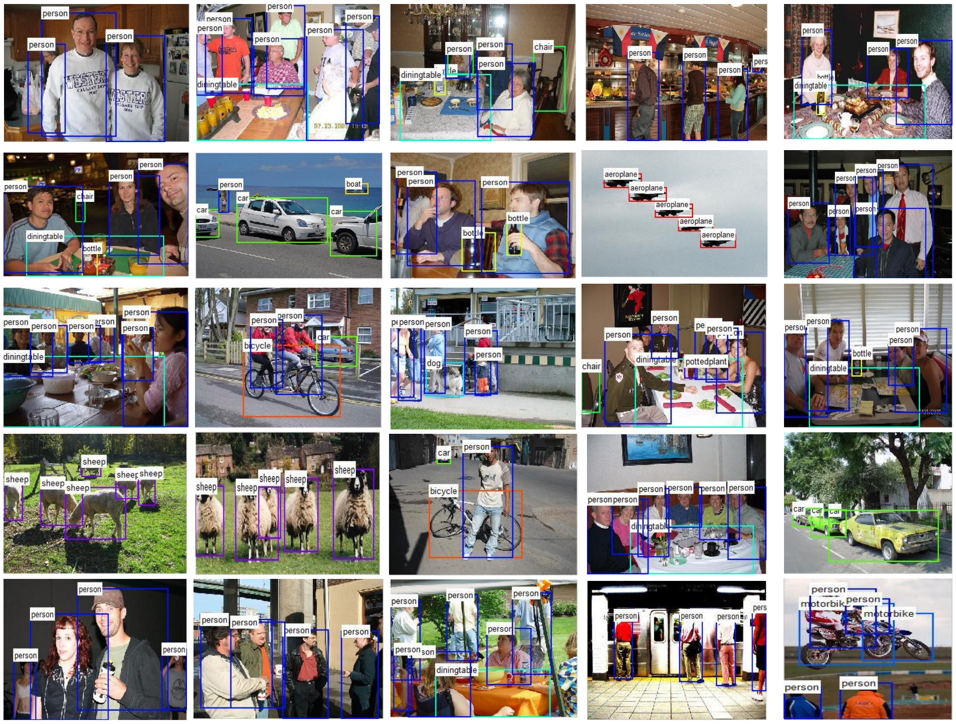

Figure 3. Example object detection results of our approach SDS validation set

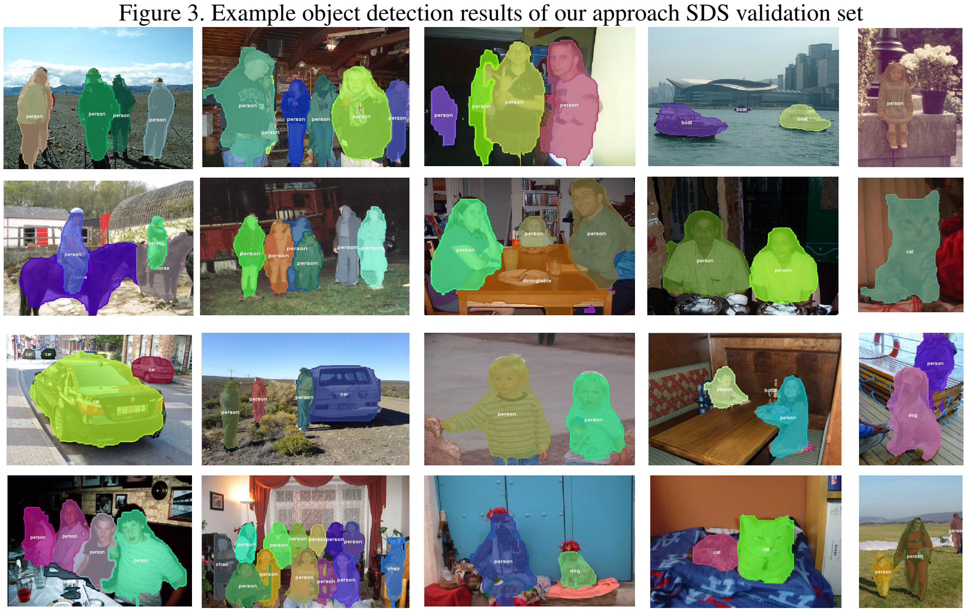

Figure 4. Example object segmentation results of our approach on SDS validation set

5. Conclusion

In this work, we present a novel Pseudo-mask Augmented object Detection (PAD) model to facilitate object detection with the instance-level segmentation information that are only supervised by bounding box annotation. Starting from the joint object detection and instance segmentation network, the proposed PAD recursively estimates the pseudo ground-truth object masks from the instance-level object segmentation network training, and then enhance the detection network with a top-down segmentation feedback.

References

[1] S. Bell, C. L. Zitnick, K. Bala, and R. Girshick. Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks. In *CVPR*, 2016.

[2] Y. Boykov and V. Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *TPAMI*, 2004.

[3] Y. Y. Boykov and M.-P. Jolly. Interactive graph cuts for optimal boundary and region segmentation of objects in nd images. In *ICCV*, 2001.

[4] X. Chen, A. Shrivastava, and A. Gupta. Enriching visual knowledge bases via object discovery and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2027–2034, 2014.

[5] J. Dai, K. He, Y. Li, S. Ren, and J. Sun. Instance-sensitive fully convolutional networks. In *ECCV*, 2016.

[6] J. Dai, K. He, and J. Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *ICCV*, 2015.

[7] J. Dai, K. He, and J. Sun. Instance-aware semantic segmentation via multi-task network cascades. In *CVPR*, 2015.

[8] J. Dai, Y. Li, K. He, and J. Sun. R-fcn: Object detection via region-based fully convolutional networks. In *NIPS*, 2016.

[9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[10] J. Dong, Q. Chen, S. Yan, and A. Yuille. Towards unified object detection and semantic segmentation. In *ECCV*. 2014.

[11] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes (VOC) Challenge. *IJCV*, 2010.

[12] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient graph-based image segmentation. *IJCV*, 2004.

[13] S. Fidler, R. Mottaghi, A. Yuille, and R. Urtasun. Bottom-up segmentation for top-down detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3294–3301, 2013.

[14] S. Gidaris and N. Komodakis. Locnet: Improving localization accuracy for object detection. *arXiv preprint arXiv:1511.07763*, 2015.

[15] S. Gidaris and N. Komodakis. Object detection via a multi-region & semantic segmentation-aware cnn model. *arXiv preprint arXiv:1505.01749*, 2015.

[16] S. Gidaris and N. Komodakis. Locnet: Improving localization accuracy for object detection. In *CVPR*, 2016.

[17] R. Girshick. Fast R-CNN. In *ICCV*, 2015.

[18] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.

[19] R. Gokberk Cinbis, J. Verbeek, and C. Schmid. Segmentation driven object detection with fisher vectors. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2968–2975, 2013.

[20] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, and J. Malik. Semantic contours from inverse detectors. In *ICCV*, 2011.

[21] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik. Simultaneous detection and segmentation. In *ECCV*. 2014.

[22] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[23] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[24] V. Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *NIPS*, 2011.

[25] T. Kong, A. Yao, Y. Chen, and F. Sun. Hypernet: Towards accurate region proposal generation and joint object detection. *arXiv preprint arXiv:1604.00600*, 2016.

[26] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the eighteenth international conference on machine learning, ICML*, volume 1, pages 282–289, 2001.

[27] Y. Li, J. Sun, C.-K. Tang, and H.-Y. Shum. Lazy snapping. In *ACM Transactions on Graphics (ToG)*, volume 23, pages 303–308. ACM, 2004.

[28] D. Lin, J. Dai, J. Jia, K. He, and J. Sun. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *CVPR*, 2016.

[29] Z. Liu, X. Li, P. Luo, C.-C. Loy, and X. Tang. Semantic image segmentation via deep parsing network. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1377–1385, 2015.

[30] M. Najibi, M. Rastegari, and L. S. Davis. G-cnn: an iterative grid based object detector. *arXiv preprint arXiv:1512.07729*, 2015.

[31] P. O. Pinheiro, R. Collobert, and P. Dollar. Learning to segment object candidates. In *NIPS*, 2015.

[32] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NIPS*, 2015.

[33] C. Rother, V. Kolmogorov, and A. Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. *ACM Transactions on Graphics*, 2004.

[34] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[35] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *European conference on computer vision*, pages 1–15. Springer, 2006.

[36] A. Shrivastava, A. Gupta, and R. Girshick. Training region-based object detectors with online hard example mining. In *CVPR*, 2016.

[37] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

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

[39] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, 2015.

[40] B. Yang, J. Yan, Z. Lei, and S. Z. Li. Craft objects from images. *arXiv preprint arXiv:1604.03239*, 2016.

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