Cyclic Guidance for Weakly Supervised Joint Detection and Segmentation

Yunhang Shen[1], Rongrong Ji[1,2]*, Yan Wang[3], Yongjian Wu[4], Liujuan Cao[1]
[1]Fujian Key Laboratory of Sensing and Computing for Smart City, School of Information Science and Engineering, Xiamen University, 361005, China, [2]Peng Cheng Laborotory, China
[3]Microsoft, Redmond, USA
[4]BestImage, Tencent Technology (Shanghai) Co.,Ltd, China
shenyunhang01@gmail.com, {rrji, caoliujuan}@xmu.edu.cn, wanyan@microsoft.com, littlekenwu@tencent.com

Abstract

*Weakly supervised learning has attracted growing research attention due to the significant saving in annotation cost for tasks that require intra-image annotations, such as object detection and semantic segmentation. To this end, existing weakly supervised object detection and semantic segmentation approaches follow an iterative label mining and model training pipeline. However, such a self-enforcement pipeline makes both tasks easy to be trapped in local minimums. In this paper, we join weakly supervised object detection and segmentation tasks with a multi-task learning scheme for the first time, which uses their respective failure patterns to complement each other's learning. Such cross-task enforcement helps both tasks to leap out of their respective local minimums. In particular, we present an efficient and effective framework termed **W**eakly **S**upervised **J**oint **D**etection and **S**egmentation (WS-JDS). WS-JDS has two branches for the above two tasks, which share the same backbone network. In the learning stage, it uses the same cyclic training paradigm but with a specific loss function such that the two branches benefit each other. Extensive experiments have been conducted on the widely-used Pascal VOC and COCO benchmarks, which demonstrate that our model has achieved competitive performance with the state-of-the-art algorithms.*

1. Introduction

In recent years, Deep Convolutional Neural Networks (DCNNs) have demonstrated outstanding capability in various computer vision tasks, such as image classification [2, 63, 37, 33], object detection [26, 54, 53, 45] and semantic segmentation [46, 13, 12, 81]. The core of the success lies in the decade-long effort to construct large-scale annotated datasets, such as ImageNet [17], PASCAL VOC [20], and COCO [43]. However, along with the great potential and flexibility in tasks like classification, the heavy dependency on the large-scale annotations had two drawbacks. First, human labeling can be expensive even with crowd sourcing. It is especially true for tasks that require pixel-level labeling such as instance segmentation. Although the community are aware that an ImageNet-like dataset for those tasks will be of great benefit, it is still absent, and may be prohibitively expensive even in the near future. Second, compared with fully annotated datasets, weakly annotated datasets (*i.e.*, only image-level labels are annotated) may be much more widely available and have much larger scale. And recent experiments show that models trained on these datasets with noisy and incomplete annotations may perform comparable or even better than models trained on fully annotated but smaller datasets [47]. Therefore, weakly supervised learn-

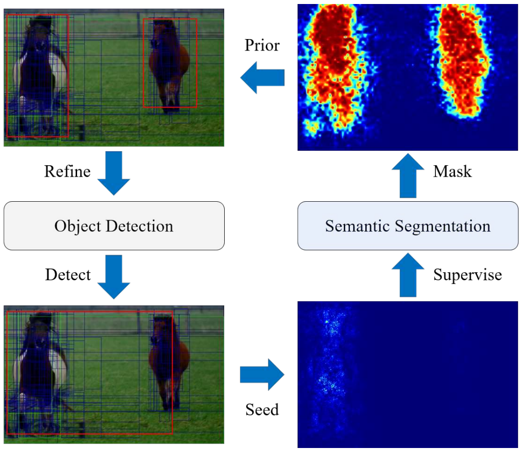

Figure 1. The core idea of the proposd cyclic guidance learning framework. Individually trained object detectors and semantic segmenters often fail on challenging cases, like the bottom-left figure shows. However, we found the failure patterns of object detection and semantic segmentation are complementary, and thus propose to train a multi-task model to allow them to benefit each other in a cyclic way. This figure is best viewed in color.

*Corresponding author.

ing has attracted growing attention [21, 67, 70, 77, 58], especially for expensive tasks in terms of annotation such as semantic segmentation and object detection, which aim at utilizing the widely available datasets in the literatures (such as PASCAL VOC and COCO) with only image-level label.

Traditional Weakly Supervised Semantic Segmentation (WSSS) and Weakly Supervised Object Detection (WSOD) are considered two separate tasks. Both tasks employ a framework of two-step iterative learning, with one step mining the labels, and the other training the model using the mined labels. An obvious problem for the framework is that the model is easy to get trapped in a local minima [15, 75, 40, 69]. Therefore, the research of WSSS and WSOD focuses on the introduction of prior and/or regularization [8, 65, 38, 60, 41, 72, 67, 66, 76, 39]. In this paper, we conquer this challenge from another aspect. Indeed, by visualizing the final as well as intermediate results of the trained models, we have some interesting observations and inspirations in the process of exploring the failure patterns of popular approaches.

As shown in Fig. 2, a WSSS neural network is often not able to obtain a label map that is consistent with object boundaries. Note how the red regions in the second column differ from the actual object of interest. This is exactly the reason why popular WSSS approaches have a graphical model such as Conditional Random Field (CRF) following the network to refine the result using additional signals [49, 50, 76, 52, 42, 34, 66, 1, 78, 1]. While CRFs demonstrate improvement on the pixel map, the quality of the final result heavily relies on the intermediate pixel map from the network. And thus aforementioned research is still mainly on the semantic labeling network.

On the other hand, to effectively take advantage of image-level annotations, WSOD approaches usually adopt a two-stage framework with traditional region proposals followed by a classification network. Different from WSSS, the presence of region proposal avoids the case that a bounding box crossing the boundary of an object most of the time. However, WSOD also has its own problems. As shown in the bottom-left subfigure of Fig. 1, a typical failure pattern for WSOD is to mis-recognize multiple objects in the same class as one single object. In some other cases, WSOD detectors would output an over-tight bounding box that only cover part of an object.

Very interestingly, the failure patterns of WSSS and WSOD are actually complementary. We argue that when tackling the problem of weakly supervised learning from image-level label, especially WSSS and WSOD, a multi-task learning framework is a necessity. On one hand, the imperfect pixel map from semantic segmentation can help the object detector leap from the local minima of over-merged or over-tight bounding boxes. On the other hand, the bounding boxes from the object detector do not have the problem of crossing object boundaries, and thus can provide a reasonably good seed for the semantic segmentation network. Actually, there are similar ideas emerging in the field of WSOD in the recent two years. For example, Wei *et al.* [77] and Diba *et al.* [19] both introduced three-stage CNNs, in which the segmentation stage leverages object localization cues from the classification stage. But the approaches did not explicitly model the mutual benefit of the object detection task and the semantic segmentation task, and thus are essentially different from the proposed approach.

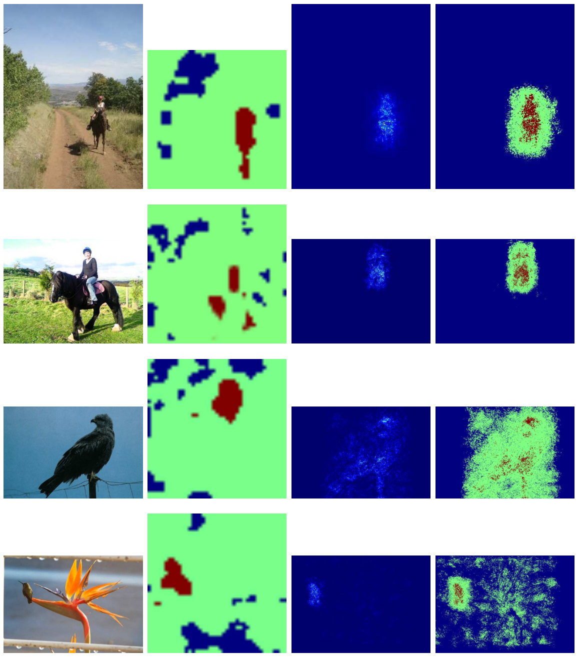
Figure 2. Comparison of different failure pattens for weakly supervised semantic segmentor and object detector. The four columns show the original images, semantation maps from CAM [85], object localization maps from our detection branch, and refined object detection maps, respectively. It is worth noting that localization map provides higher quality background cue than the classification maps. The right three columns are drawn with jet color scale, where red color corresponds to high value and blue color corresponds to low value. This figure is best viewed in color.

In this paper, we present a Weakly Supervised Joint Detection and Segmentation (WS-JDS) framework. The core is a backbone deep network supporting two branches for object detection and semantic segmentation, respectively. Regarding to model training, we propose a Cyclic Guidance Learning (CGL) approach, as illustrated in Fig. 1. Similar with traditional WSSS and WSOD approaches, CGL iteratively does label mining and model training. But when training the object detection branch, we use the bounding boxes derived from both the filtered region proposal and the segmentation pixel map as training data. And the localization cues from the object detection branch are also used to supervise the training of the semantic segmentation branch.

To demonstrate the effectiveness of the proposed network as well as the training scheme, we present detailed

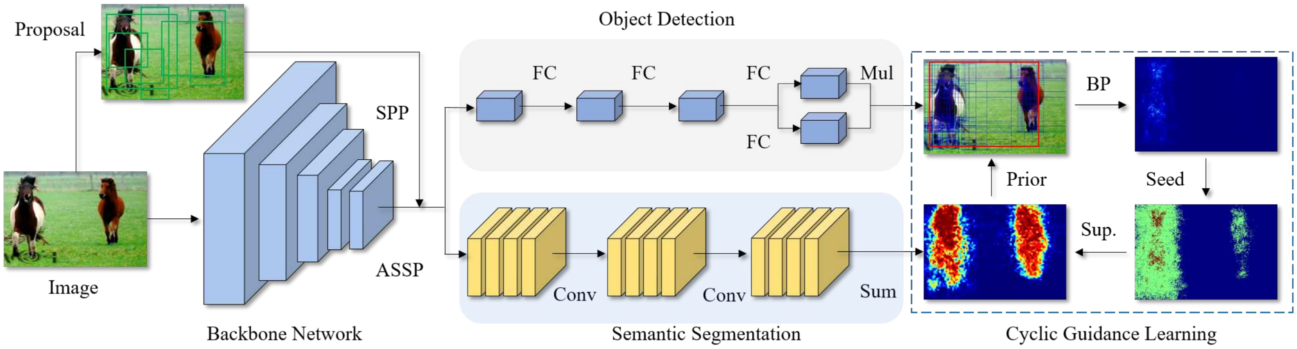

Figure 3. Overview of CGL for WS-JDS framework. Typical CNN layers are leveraged to extract the intermediate features of an input image as the backbone. In detection branch, features for each object proposal are generated by Spatial Pyramid Pooling (SPP) layer followed by two fully connected layers. A two stream detector [10] is utilized to discovery object instances under image-level supervision. The object localization map is extracted through gradient-based saliency method via Back-Propagation (BP). In segmentation branch, the entire feature map are firstly fed into a fully convolutional sub-network to predict segmentation mask via Atrous Spatial Pyramid Pooling (ASPP), and then supervised (Sup.) by the rough segmentation masks produced by object localization map. Meanwhile, based on the segmentation mask, we evaluate the proposals confidence of contained object instance properly.

evaluations on both tasks of object detection and semantic segmentation. The evaluation of object detection is performed on PASCAL VOC 2007, 2010 and 2012 [20], with comparison with several state-of-the-art methods [23, 70, 68, 77, 72, 82, 24, 58]. We also evaluate our method on COCO dataset [43] for object detection and the instance segmentation tasks. On both tasks, we demonstrate competitive performance with the state-of-the-art methods.

2. Related Work

Weakly Supervised Object Detection. WSOD refers to learning an object detector with only image-level annotations that indicate the presence of an category. Based on the optimization objective, WSOD methods can be divided into two groups, *i.e.*, object discovery and instance refinement.

The object discovery approaches optimize the image-level classification loss based on traditional object proposal directly, *i.e.*, formulate the WSOD problem with a Multi-Instance Classification (MIC) paradigm. The learning step of MIC alternates between selecting positive samples and training an appearance model. A number of different strategies to train the MIC model had been proposed in the literature [15, 73, 9, 75, 60, 23, 57]. Recent approaches combined Convolutional Neural Networks (CNNs) and MIC into a unified framework [10, 19]. Contextual information was introduced to achieve promising improvement [38]. Recently, Tang *at al.* [70] proposed to replace traditional object proposal extraction stage by generating and refining object proposals in an end-to-end framework. There are some methods focused on proposal-free paradigms by taking advantage of deep feature map [7, 5, 87] and class activation maps [85, 28, 83]. However, this paradigm seriously depends on the quality of feature maps and is hard to distinguish different instances in challenging scenes. Some work also used additional annotations and data to improve the performance, *e.g.*, object size estimation [60], instance count annotation [22], video motion cue [64] and human verification [48]. Some of the additional data may be from a different domain. Therefore, knowledge transfer for progressive cross-domain adaptation was also exploited, *e.g.*, data domain adaption [59] and task domain adaption [35].

The instance refinement approaches also follow the bounding box mining and model training framework. But instead of optimizing a MIC loss, they optimize the objective function of instance-level localization. Therefore, from another prospect, they have an additional instance refinement stage after object discovery by introducing a fully supervised detector. For example, the work in [40, 36, 22, 69] mined the high-confidence proposals and treated them as positive samples to train a fully supervised model. Many efforts [84, 24] had been made to mining high-quality bounding boxes. To further improve the robustness, there are some works that combined the weakly supervised MIC model and fully supervised detectors. For example, Tang *et al.* [69] introduced multiple supervised branches to refine the result from weakly supervised model. Work in [41, 72] proposed min-entropy prior to alleviate the ambiguity of result and used the pseudo ground-truth object to optimize the objective function of localization. Zhang *et al.* [82] proposed to estimate sample-wise training difficulty to learn a fully supervised detector in an easy-to-difficult order.

Weakly Supervised Semantic Segmentation. WSSS methods can also been divided into two groups. The first group [50, 61, 55, 86] leverages CNN built-in pixel-level cues and constraint priors to learn segmentation masks, while a common practice for the second group [39, 78, 1, 34] is to treat initial object localization cues, (which is often produced by classification networks,) as pseudo supervision and train a fully supervised segmentation network.

In the first group, Pathak *et al.* [50] proposed a constrained CNN, which applied linear constraints on the structured output space of pixel labels. Saleh *et al.* [55] extracted the built-in masks directly from the hidden layer activation, and incorporated the resulting masks via a weakly supervised loss. There are also works that derive category-wise saliency maps from intermediate feature maps of CNNs to estimate the segmentation masks [61, 86].

In the second group, popular methods [51, 11, 21] leveraged object saliency map or feature activation map to provide complimentary information. Many priors or regularization [39, 66, 67] were proposed to improve the segmentation result. Different kinds of supervisions were exploited: web data [32, 56], bounding boxes [79], scribbles [42], points [6], *etc*. There are also work [76, 52] that focused on improving feature learning in iterative frameworks. Recently, various approaches based on iteratively mining common feature [74], seeded region growing [34], random-walk label propagation [71], dilated convolution [78] and pixel-level semantic affinity [1] were proposed.

Multi-task Learning. Learning detection and segmentation jointly was first employed by Hariharan *et al.* [29] in fully supervised learning. Although the framework in [29] was multi-stages, it still showed improvement of performance on individual task. He *et al.* [30] also demonstrated that box detection can benefit from multi-task learning. Recent works provided more complex mechanism to combine the two tasks with the assistance of direction prediction [12], and information flow boosting [44]. In weakly supervised learning, some related work used segmentation masks to boost the performance of detection task [19, 77]. However, different to those works, our method also exploits to improve segmentation branch with detection result via CGL. And in weakly supervised setting, ours is the first to join object detection and semantic segmentation tasks.

3. The Proposed Method

Overview: The overall architecture of the proposed approach is illustrated in Fig. 3. Sharing the same backbone, which is VGG16 [63], the proposed model has two branches, *i.e.*, object detection and semantic segmentation. In particular, the object detection branch, built on top of spatial pyramid pooling layer, produces box prediction and object localization map. Following the previous weakly supervised semantic segmentation approaches [78, 39, 55], we leverage the inferred localization maps to produce pseudo-ground-truth of segmentation masks from training images, which are then used as supervision to train the segmentation branch. The predicted confidence masks from the segmentation branch are then employed to evaluate object proposals on the likelihood of containing the object instance, which in turn benefits the object detection branch.

Object Detection Branch: We employ WSDDN [10] for the object detection branch and further improve the performance using the CGL scheme. In particular, let $I \in \Re^{H \times W \times 3}$ be an input image, $\mathbf{t} \in \{0,1\}^C$ be the corresponding image-level labels, and C be the total number of categories. H and W are the image height and width, respectively. As illustrated by the gray region in Fig. 3, we first extract feature of R object proposals $\{p_1 \dots p_R\}$ from the VGG backbone, and then the feature from the Spatial Pyramid Pooling (SPP) layer [31] is forked into two streams, *i.e.*, classification stream and detection stream, producing two score matrices $X^c, X^d \in \mathbb{R}^{R \times C}$ by two fully-connected layers, respectively. Both score matrices are normalized by softmax functions $\sigma(\cdot)$ over categories and proposals, respectively. Then the element-wise product of the output of the two streams is again a score matrix: $X^s = \sigma(X^c) \odot \sigma(X^d)$. To acquire image-level classification scores, a sum pooling is applied: $\mathbf{y}_k = \sum_{r=1}^{R} X^s_{rk}$. Then we obtain a cross-entropy loss function $\mathcal{L}_{\text{det}}$:

$$\mathcal{L}_{\text{det}} = \sum_{k=1}^{C} \Big\{ \mathbf{t}_k \log \mathbf{y}_k + (1 - \mathbf{t}_k) \log(1 - \mathbf{y}_k) \Big\}, \qquad (1)$$

where $\mathbf{t}_k$ is the ground truth labels of whether an object of category k is presented in the image I.

Prior Guidance: However, such object discovery optimization lacks prior guidance. Note X^s is calculated based on local information of each individual proposal [58]. The correlation among instances is typically ignored and the optimization might converge to an undesirable local minimum during MIC learning [3]. Recent work in [38, 77] proposed to use contextual information as a supervisory guidance, which enforces the predicted object region to be compatible with its surrounding context. We propose to leverage knowledge of the learned masks from segmentation branch to refine the detection via objectness prior [77, 58].

Semantic Segmentation Branch: In order to obtain the segmentation masks, we first collect the intermediate features before the last pooling layer of the backbone network. Then we feed it to the convolutional blocks with multiple dilated rates to localize object-related regions perceived by different receptive fields, similar to DeepLab-ASPP [14], as illustrated by the blue region in Fig. 3. With the produced object localization cues, we train the segmentation branch with pixel-wise loss $\mathcal{L}_{\text{seg}}$, which is widely adopted by fully supervised schemes. Different from previous literature [39, 55, 14, 78] of fully/weakly supervised semantic segmentation, when applying Fully Convolutional Networks (FCNs) to semantic segmentation, which typically uses a per-pixel softmax and a multinomial cross-entropy loss, we use a per-pixel Sigmoid and a binary cross-entropy loss. Then $\mathcal{L}_{\text{seg}}$ is similar to $\mathcal{L}_{\text{det}}$, but with additional spatial dimensions:

$$\mathcal{L}_{\text{seg}} = \sum_{k,h,w}^{C,\hat{H},\hat{W}} \Big\{ M^k_{hw} \log S^k_{hw} + (1 - M^k_{hw}) \log(1 - S^k_{hw}) \Big\}, \qquad (2)$$

where M and S denote the rough segmentation masks produced by detection branch and the predicted masks with C

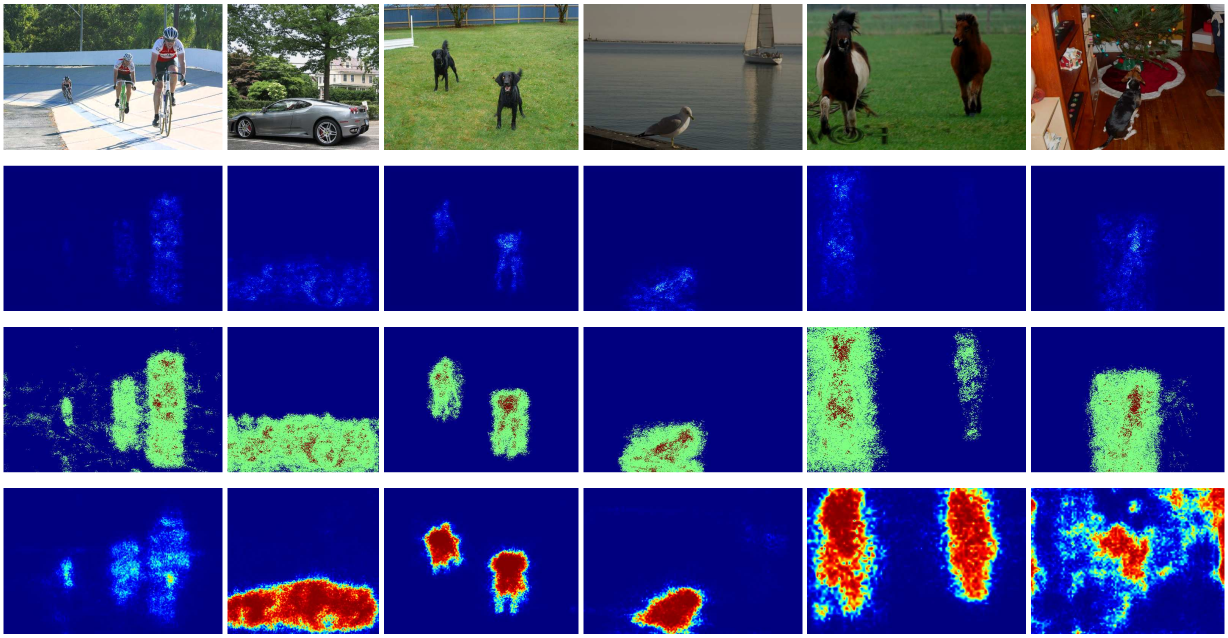

Figure 4. Visualization of the intermediate steps of the proposed CGL. The rows show input image, object localization map from object detection branch, rough segmentation map derived from the localization map, and output of segmentation branch without CRF post-processing, respectively. Red, blue, and green pixels in the third row indicate foreground, background and uncertain, respectively.

channels, respectively. And $\hat{H}, \hat{W}$ denote the image height and width of the predicted masks, which is usually $1/16$ of H, W. Besides, we also use a constrain-to-boundary loss from [39] to encourage segmentation masks to match up with object boundaries.

Our definition of $\mathcal{L}_{\text{seg}}$ allows the segmentation network to generate masks for every category without competition among categories. We rely on the dedicated detection branch to predict the category label used to select the output masks. As demonstrated in [30], by using this decoupled mask and category prediction, once the instance has been classified as a whole (by the detection branch), it is sufficient to predict binary masks without concern for the categories, which makes the model easier to train.

Cyclic Guidance Learning: Theoretically, the loss functions of WSOD and WSSS lead to complementary failure patterns. On one hand, most works formulate the WSOD problem with an MIC paradigm. Its explicit penalty on false positives from negative bags gives WSOD low false positive rate. However, to prevent self-reinforcing into a local minimum, popular loss only penalizes confident false negatives (which gives limited pseudo-ground-truth) with an IoU less than a threshold (which compromises sensitivity). Consequently, WSOD usually suffers from ambiguous feature maps around non-discriminative parts of objects. On the other hand, for WSSS, the loss is defined on pixel-level. The lack of explicit penalty on false positives often results in noisy background. But the fine granularity gives better it precision on ambiguous regions to guide object localizer.

We propose a CGL scheme to exploit complementary knowledge learned by individual tasks, as illustrated by the blue dashed line in Fig. 3. For the detection-to-segmentation guidance, we leverage the inferred localization maps to produce rough segmentation masks M in Eq. 2, which are then used as supervision to train the segmentation branch. Different from [19, 77] that introduced extra saliency detection and classification branches to generate localization maps [85, 50], we produce built-in background and foreground cues from the detection branch through gradient-based saliency detection following [62, 39], which has the benefit of parameter free. In particular, the gradient of classification score flows from detection branch to the first layer of backbone by back-propagation, which is illustrated in the second row of Fig. 4. On the object localization maps, we assign the pixels with values larger than a pre-defined normalized threshold (*i.e.*, 0.1) with the corresponding category label as the foreground regions. We also choose pixels with low normalized value (*i.e.*, 0.005) as background sample. The remaining pixels are marked as uncertain and ignored during training. The result foreground, background and uncertain pixels from sample images are illustrated in the third row of Fig. 4.

The last row of Fig. 4 illustrates that the output of segmentation branch is able to generalize object localization seed to predict uncertain pixels, which in turn provides guidance to the detector training. For example, a false positive detection occurs in the fifth column, when the detection branch fails to discover multiple instance of category *horse* existed in the image. And therefore object localization map is half-baked (second row). In this case, the image-level annotation cannot correct this problem, which ends up with a pseudo ground truth of segmentation with most

Table 1. Object detection on PASCAL VOC 2007 in terms of AP (%) on test set.

Method	aero	bicy	bird	boa	bot	bus	car	cat	cha	cow	dtab	dog	hors	mbik	pers	plnt	she	sofa	trai	tv	Av.
WSDDN [10]	39.4	50.1	31.5	16.3	12.6	64.5	42.8	42.6	10.1	35.7	24.9	38.2	34.4	55.6	9.4	14.7	30.2	40.7	54.7	46.9	34.8
WSDDN*	45.9	48.1	32.4	13.3	23.0	61.7	51.1	40.7	16.8	37.9	23.8	28.4	43.1	53.0	6.5	21.1	41.2	44.0	60.6	45.9	36.9
WSOD $\|\|$ WSSS	50.1	56.3	32.4	22.7	19.0	51.8	41.1	62.6	2.7	45.3	45.6	24.4	43.7	56.0	12.4	20.5	38.1	34.8	53.2	33.4	37.3
WSOD $\rightarrow$ WSSS	52.4	63.5	28.8	16.1	27.3	58.0	55.8	41.6	22.5	47.3	14.0	25.9	8.5	55.2	18.9	22.1	46.9	45.0	54.8	49.9	37.7
WSOD $\leftarrow$ WSSS	39.8	61.4	34.6	18.1	27.3	66.1	52.9	50.8	15.6	43.0	42.4	46.1	19.4	57.9	30.5	24.2	44.0	48.2	64.8	52.9	42.0
WSOD $\rightleftharpoons$ WSSS	52.0	64.5	45.5	26.7	27.9	60.5	47.8	59.7	13.0	50.4	46.4	56.3	49.6	60.7	25.4	28.2	50.0	51.4	66.5	29.7	45.6

pixels marked as background for the second instance (third row). However, the segmentation branch is able to predict a coarse mask properly to overturn the mistake (last row). Thus the segmentation map provides supervisor guidance to refine the detector. Another example of object detection branch benefits segmentation branch is illustrated in the last column of Fig. 4, when the segmentation branch fails to predict coarse mask properly of *dog* existed in the image, the detection branch provides conservative seed of *dog* and clear supervision of background.

For the segmentation-to-detection guidance, we treat the learned masks as localization prior to refine the proposal classification following [58]. In particular, the masks provide contextual information for each proposal. Given the r-th proposal and the k-th category, the confident weight is estimated from the masks S:

$$W_{rk} = \frac{1}{\sqrt{|p_r|}} \sum_{i,j \in p_r} T(S_{ij}^k) - \frac{1}{\sqrt{|p_r^c|}} \sum_{i,j \in p_r^c} T(S_{ij}^k), \quad (3)$$

where $T(S_{ij}^k) = \mathbb{1}[S_{ij}^k \geq 10^{-1} \cdot \max S^k]$, p_r and p_r^c is the r-th proposal and the corresponding contextual region. The contextual region p_r is defined as the surrounding regions of p_r by scaling the box by a factor of 1.8 [25]. Therefore, before computing image-level score with sum pooling, we refine the predicted proposal score X^s with W by element-wise (Hadamard) product. And we get the refined cross-entropy loss function as:

$$\mathcal{L}_{\text{det}}^{\text{r}} = \sum_{k=1}^{C} \Big\{ \mathbf{t}_k \log \mathbf{y}_k^{\text{r}} + (1 - \mathbf{t}_k) \log(1 - \mathbf{y}_k^{\text{r}}) \Big\}, \quad (4)$$

where $\mathbf{y}_k^{\text{r}} = \sum_{r=1}^{R} W_{rk} X_{rk}^s$. However, mask S is unstable in early training iterations. Therefore, we also use object localization map to refine the proposal classification.

During the testing stage, we run the box prediction branch on these proposals, followed by a non-maximum suppression. At the same time, the mask prediction branch outputs the segmentation masks for the entire image. Then we extract the masks of detection boxes.

4. Experimental Evaluation

4.1. Datasets and Evaluation Protocol

Dataset. We evaluate the proposed approach on PASCAL VOC 2007, 2010, 2012 [20] and COCO [43], which are widely-used benchmark datasets. PASCAL VOC 2007 consists of $2,501$ training images, $2,510$ validation images, and $4,092$ test images over 20 categories. PASCAL VOC 2010 consists of $4,998$ training images, $5,105$ validation images, and $9,637$ test images over 20 categories. PASCAL VOC 2012 consists of $5,717$ training images, $5,823$ validation images, and $10,991$ test images over 20 categories. Following the standard settings of weakly supervised object detection, we use both training and validation sets with only image-level labels for training. The performance of the localization task, defined as predicting boxes when categories are known, is evaluated on the training and validation sets, and the performance of the detection task, defined as predicting categories and boxes simultaneously, is evaluated on the testing set, following common practice [18, 10]. Note our evaluation settings are more challenging than some popular approaches, such as [15, 8, 9], which removed the hard images containing only truncated and difficult objects. We also evaluate our approach on the MS COCO dataset [43], which is among the most challenging datasets for instance segmentation and object detection. It consists of 80 object categories with pixel-wise instance mask annotations. Our experiments involve the 115k training set, 5k validation set. Only image-level annotations are used in training.

Evaluation Protocols. Two protocols are used for evaluation: CorLoc and mean Average Precision (mAP). CorLoc is a commonly used measurement that quantifies localization performance by the percentage of images that contain at least one object instance with at least 50% overlapped to the ground-truth. CorLoc indicates the ratio of images in which a method correctly localizes an object of the target category according to the PASCAL-criterion. The mAP follows standard PASCAL VOC protocol to report the mAP at 50% Intersection-over-Union (IoU) of the detected boxes with the ground-truth. We evaluate the CorLoc and mAP on the training/validation and testing splits, respectively. For MS COCO data, we also report the standard COCO metrics including AP (averaged over IoU thresholds), AP$_{50}$, AP$_{75}$, AP$_S$, AP$_M$, and AP$_L$ (AP at different scales). We use the superscripts r and b for object detection AP and instance segmentation AP, respectively.

4.2. Implementation Details

The proposed approach is implemented using Caffe2. Both Python and C++ interfaces are used. For the backbone network, we use VGG16 [63] that is initialized with the weights pretrained on ImageNet [17]. We use WSDDN [10] as our baseline model for the WSOD branch.

Table 2. Object detection on PASCAL VOC 2007 in terms of AP (%) on test set.

Method		aero	bicy	bird	boa	bot	bus	car	cat	cha	cow	dtab	dog	hors	mbik	pers	plnt	she	sofa	trai	tv	Av.
WSDDN VGG16	[10]	39.4	50.1	31.5	16.3	12.6	64.5	42.8	42.6	10.1	35.7	24.9	38.2	34.4	55.6	9.4	14.7	30.2	40.7	54.7	46.9	34.8
WCCN	[19]	49.5	60.6	38.6	**29.2**	16.2	**70.8**	56.9	42.5	10.9	44.1	29.9	42.2	47.9	64.1	13.8	23.5	45.9	**54.1**	60.8	54.5	42.8
Jie *et al.*	[36]	52.2	47.1	35.0	26.7	15.4	61.3	66.0	54.3	3.0	53.6	24.7	43.6	48.4	65.8	6.6	18.8	**51.9**	43.6	53.6	62.4	41.7
OICR-VGG16	[69]	**58.0**	62.4	31.1	19.4	13.0	65.1	62.2	28.4	**24.8**	44.7	30.6	25.3	37.8	65.5	15.7	24.1	41.7	46.9	64.3	**62.6**	41.2
SPAM-CAM	[28]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	27.5
TST	[59]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	33.8
TS^{2}C	[77]	59.3	57.5	43.7	27.3	13.5	63.9	61.7	**59.9**	24.1	46.9	36.7	45.6	39.9	62.6	10.3	23.6	41.7	52.4	58.7	56.6	44.3
Ge *et al.*	[23]	49.1	53.6	43.5	21.3	18.5	66.9	64.0	55.6	11.9	53.7	26.6	45.6	48.7	64.6	20.4	23.3	50.0	44.7	55.9	60.6	43.9
Tang *et al.*	[70]	57.9	**70.5**	37.8	5.7	21.0	66.1	**69.2**	59.4	3.4	**57.1**	**57.3**	35.2	**64.2**	**68.6**	**32.8**	**28.6**	50.8	49.5	41.1	30.0	45.3
WS-JDS		52.0	64.5	**45.5**	26.7	**27.9**	60.5	47.8	59.7	13.0	50.4	46.4	**56.3**	49.6	60.7	25.4	28.2	50.0	51.4	**66.5**	29.7	**45.6**
OICR FRCNN	[69]	65.5	67.2	47.2	21.6	22.1	68.0	68.5	35.9	5.7	63.1	49.5	30.3	64.7	66.1	13.0	25.6	50.0	57.1	60.2	59.0	47.0
MELM	[72]	55.6	66.9	34.2	29.1	16.4	68.8	68.1	43.0	**25.0**	65.6	45.3	53.2	49.6	68.6	2.0	25.4	52.5	56.8	62.1	57.1	47.3
ZLDN	[82]	55.4	68.5	50.1	16.8	20.8	62.7	66.8	56.5	2.1	57.8	47.5	40.1	69.7	68.2	21.6	27.2	53.4	56.1	52.5	58.2	47.6
Ge *et al.*	[24]	64.3	68.0	**56.2**	**36.4**	23.1	68.5	67.2	64.9	7.1	54.1	47.0	57.0	69.3	65.4	20.8	23.2	50.7	59.6	65.2	57.0	51.2
W2F	[84]	63.5		50.5	31.9	14.4	72.0	67.8	**73.7**	23.3	53.4	49.4	**65.9**	57.2	67.2	27.6	23.8	51.8	58.7	64.0	**62.3**	52.4
TS^{2}C FRCNN	[77]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	48.0
Tang *et al.* FRCNN	[70]	63.0	69.7	40.8	11.6	27.7	70.5	**74.1**	58.5	10.0	66.7	**60.6**	34.7	**75.7**	70.3	25.7	26.5	55.4	56.4	55.5	54.9	50.4
WS-JDS FRCNN		**64.8**	**70.7**	51.5	25.1	**29.0**	**74.1**	69.7	69.6	12.7	**69.5**	43.9	54.9	39.3	**71.3**	**32.6**	**29.8**	**57.0**	**61.0**	**66.6**	57.4	**52.5**

Table 3. Object localization on PASCAL VOC 2007 in terms of CorLoc (%) on trainval set.

Method		aero	bicy	bird	boa	bot	bus	car	cat	cha	cow	dtab	dog	hors	mbik	pers	plnt	she	sofa	trai	tv	Av.
WSDDN VGG16	[10]	65.1	58.8	58.5	33.1	39.8	68.3	60.2	59.6	34.8	64.5	30.5	43.0	56.8	82.4	25.5	41.6	61.5	55.9	65.9	63.7	53.5
WCCN VGG16	[19]	83.9	72.8	64.5	44.1	40.1	65.7	82.5	58.9	33.7	72.5	25.6	53.7	67.4	77.4	26.8	49.1	68.1	27.9	64.5	55.7	56.7
Jie *et al.*	[36]	72.7	55.3	53.0	27.8	35.2	68.6	81.9	60.7	11.6	71.6	29.7	54.3	64.3	88.2	22.2	53.7	72.2	52.6	68.9	75.5	56.1
OICR-VGG16	[69]	81.7	80.4	48.7	**49.5**	32.8	**81.7**	85.4	40.1	40.6	79.5	35.7	33.7	60.5	88.8	21.8	57.9	76.3	59.9	75.3	**81.4**	60.6
SP-VGGNet	[87]	**85.3**	64.2	67.0	42.0	16.4	71.0	64.7	**88.7**	20.7	63.8	58.0	**84.1**	**84.7**	80.0	**60.0**	29.4	56.3	**68.1**	77.4	30.5	60.6
TST	[59]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	59.5
TS^{2}C	[77]	84.2	74.1	61.3	52.1	32.1	76.7	82.9	66.6	**42.3**	70.6	39.5	57.0	61.2	88.4	9.3	54.6	72.2	60.0	65.0	70.3	61.0
Ge *et al.*	[23]	75.9	67.6	62.2	37.3	36.6	71.5	80.2	63.8	19.7	70.6	32.4	56.1	67.8	81.7	35.9	50.9	73.4	50.4	66.0	66.8	58.3
Tang *et al.*	[70]	77.5	**81.2**	55.3	19.7	44.3	80.2	**86.6**	69.5	10.1	**87.7**	**68.4**	52.1	84.4	**91.6**	57.4	**63.4**	**77.3**	58.1	57.0	53.8	63.8
WS-JDS		82.9	74.0	**73.4**	47.1	**60.9**	80.4	77.5	78.8	18.6	70.0	56.7	67.0	64.5	84.0	47.0	50.1	71.9	57.6	**83.3**	43.5	**64.5**
OICR FRCNN	[69]	**85.8**	82.7	62.8	**45.2**	43.5	**84.8**	87.0	46.8	15.7	82.2	51.0	45.6	83.7	91.2	22.2	59.7	75.3	65.1	76.8	**78.1**	64.3
ZLDN	[82]	74.0	77.8	65.2	37.0	46.7	75.8	83.7	58.8	17.5	73.1	49.0	51.3	76.7	87.4	30.6	47.8	75.0	62.5	64.8	68.8	61.2
W2F	[84]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	70.3
Tang *et al.* FRCNN	[70]	83.8	82.7	60.7	35.1	53.8	82.7	**88.6**	67.4	22.0	**86.3**	**68.8**	50.9	**90.8**	**93.6**	44.0	**61.2**	**82.5**	65.9	71.1	76.7	68.4
WS-JDS FRCNN		79.8	**84.0**	**68.3**	40.2	**61.5**	80.5	85.8	**75.8**	**29.7**	77.7	49.5	**67.4**	58.6	87.4	**66.2**	46.6	78.5	**73.7**	**84.5**	72.8	**68.6**

Table 4. Object detection and localization on PASCAL VOC 2010 and 2012 in terms of mAP (%) and CorLoc (%).

Method		2010		2012	
		mAP (%)	CorLoc (%)	mAP (%)	CorLoc (%)
Multi-Fold MIL	[16]	27.4	55.2	–	–
OICR-VGG16	[69]	–	–	37.9	62.1
Jie *et al.*	[36]	–	–	38.3	58.8
TS^{2}C	[77]	–	–	40.0	64.4
Tang *et al.*	[70]	–	–	**40.8**	**64.9**
WS-JDS		**39.9**	**63.1**	39.1	63.5
OICR FRCNN	[69]	–	–	42.5	65.6
MELM	[72]	–	–	42.4	–
ZLDN	[82]	–	–	42.9	61.5
W2F	[84]	–	–	**47.8**	69.4
TS^{2}C FRCNN	[77]	–	–	44.4	–
Tang *et al.* FRCNN	[70]	–	–	45.7	69.3
WS-JDS FRCNN		**45.7**	**68.1**	46.1	**69.5**

Training. We use a mini-batch size of 128, learning rate of 0.001, momentum of 0.9, and dropout rate of 0.5. We use a step learning rate decay schema with decay weight $\gamma = 0.1$ and step size of 20 epochs. In the multi-scale setting, we use five scales $\{480, 576, 688, 864, 1200\}$. To improve the robustness, we randomly adjust the exposure and saturation of the images by up to a factor of 1.5 in the HSV space. And a random crop with 0.9 of the original images size is applied. We use MCG [4] to generate object proposals for all experiments, including our implementation of baseline methods. We set the max number of region proposals in an image to be $2,048$. All models are trained for 30 epochs. We apply Xavier [27] and Gaussian initialization to the new convolutional and fully-connected layers, respectively.

Testing. The learned detectors are evaluated in two paradigms, following [40, 19, 69, 24, 77, 58, 70]: The first paradigm directly applies the learned detectors on the testing images and outputs the scores for each region proposal as the detection results. The second paradigm labels bounding boxes in training/validation images using WSOD, and then uses these bounding boxes as pseudo ground-truth to train the fully supervised detector, which is Fast-RCNN [26] in our case, for testing. In this scenario, for each category, we treat the proposal with maximum detection score as the pseudo ground-truth bounding box. The test scores are the average of all scales and flips. Detection results are post-processed by non-maximum suppression using a threshold of 0.5 IoU. The predicted masks are upsampled to match the size of the input image, and then apply a fully-connected CRF to refine the result.

4.3. Comparison to Baselines

To demonstrate the necessity and benefit of learning WSSS and WSOD models simultaneously, we compare our full framework with baseline models with different designs removed in Tab. 1. The first variation (*WSOD* $||$ *WSSS*) trains two tasks independently with the shared backbone network. The second (*WSOD* $\rightarrow$ *WSSS*) and third (*WSOD* $\leftarrow$ *WSSS*) variations employ only one direction guidance, *i.e.*, the detection-to-segmentation or the segmentation-to-detection guidance, respectively. The fourth (*WSOD* $\rightleftharpoons$

Table 5. Instance segmentation and object detection on minival set of COCO.

Method		AP^r	AP$^r_{50}$	AP$^r_{75}$	AP^r_S	AP^r_M	AP^r_L	AP^b	AP$^b_{50}$	AP$^b_{75}$	AP^b_S	AP^b_M	AP^b_L
WSDDN*	BB	3.4	9.5	2.9	1.0	3.7	9.2	9.5	19.2	8.2	2.1	10.4	17.2
	ELL	4.5	10.3	4.3	1.3	4.3	9.1						
	MCG	5.2	10.7	5.1	1.8	6.3	12.0						
ContextLocNet*	BB	4.2	10.3	4.1	2.1	5.3	10.1	9.9	19.4	8.7	2.1	10.8	17.9
	ELL	4.7	10.6	4.4	1.3	5.4	10.0						
	MCG	5.5	10.9	5.3	**2.0**	6.7	11.9						
WS-JDS		**6.1**	**11.7**	**5.5**	1.5	**7.1**	**12.2**	**10.5**	**20.3**	**9.2**	**2.2**	**10.9**	**18.3**

WSSS) one is our full CGL scheme. The performance of WSOD is only slightly improved in *WSOD* $||$ *WSSS* and *WSOD* $\rightarrow$ *WSSS* compared with WSDDN, which is mainly because that the WSOD model is trained independently without guidance from WSSS in these two baselines. The performance boost mainly benefits from sharing the backbone network of multi-task learning. In *WSOD* $\leftarrow$ *WSSS*, the performance of WSOD is significantly improved by exploiting segmentation pixel maps to refine the mined supervision. Finally, the proposed CGL scheme (*WSOD* $\rightleftharpoons$ *WSSS*) further improves the performance by combining guidance from both directions. As comparing to WS-DDN [10], Tab. 1 shows that our model reaches 45.6% mAP for weakly supervised object detection. Although our reproduction of WSDDN on VGG16 backbone (WSDDN*) is superior to the original WSDDN, our method still outperforms this baseline with a large margin. Experimental results demonstrate that the complementary knowledge of detection and segmentation can benefit individual training.

4.4. Comparison to the State of the Arts

PASCAL VOC. We divide the compared WSOD methods into two categories: object discovery and instance refinement based methods, as mentioned in Sec. 2 and in the first and second parts in Tab. 2 3. For fair comparison, we do not include methods that use additional data [60, 64, 22]. For object discovery, we compare our method with the state of the arts, including ZLDN [82], MELM [72], TS^{2}C [77], OICR [69] among others. The proposed model reaches 45.6% mAP, and achieves state-of-the-art performance. It is worth noting that the improvement from our framework is orthogonal to those works, so the proposed CGL framework can also benefit from all the techniques proposed in the aforementioned literatures. For instance refinement, we also train a Fast-RCNN [26] with the pseudo ground-truth localization extracted from our weakly supervised detectors. We achieve a performance of 52.5% mAP, which is superior to previous work in [40, 80, 69, 72, 82, 58, 77, 70, 84] with gain of about $0.1 \sim 5.5\%$ in Tab. 2. We further conduct experiments on PASCAL VOC 2010 and 2012. Tab. 4 shows our method consistently achieves competitive performance to the state-of-the-art approaches on all metrics.

COCO. With the proposed technique, we perform instance segmentation on the COCO, which is more challenging than the PASCAL VOC. To the best of our knowledge, this is the first work reporting results for image-level supervised instance segmentation on COCO. We construct several baselines based on object bounding boxes obtained from weakly supervised localization methods following [86]. We use three mask extraction strategies: The first strategy uses the entire bounding boxes as the instance masks (BB). The second strategy fits a maximum ellipse on the bounding boxes (ELL). The third strategy retrieves a max overlap segmentation mask in MCG with the bounding boxes (MCG). As illustrated in Tab. 5, our method achieves better performance in termed of AP^r compared with all other methods in the instance segmentation task. We also report performance of the object detection task on COCO. The proposed approach outperforms the baselines methods by 1.0% and 0.6% in AP^b, respectively.

5. Conclusion

In this paper, we propose a multi-task learning framework for the problems of weakly supervised object detection and semantic segmentation. We found that the different failure patterns of the two tasks can actually benefit each other and alleviate the problem of the optimization getting stuck in local minimum. To leverage the complementary knowledge learned by the two tasks, we further propose a Cyclic Guidance Learning scheme. In this scheme, the detection branch provides a reasonably good seed for segmentation branch, while the learned masks help the detector to leap from local minimum. On the widely-used benchmarks of Pascal VOC and COCO, the proposed method achieves competitive or superior performance to state-of-the-art methods in both weakly supervised object detection and instance segmentation tasks.

6. Acknowledgment

This work is supported by the National Key R&D Program (No.2017YFC0113000, and No.2016YFB1001503), Nature Science Foundation of China (No.U1705262, No.61772443, and No.61572410), Post Doctoral Innovative Talent Support Program under Grant BX201600094, China Post-Doctoral Science Foundation under Grant 2017M612134, Scientific Research Project of National Language Committee of China (Grant No. YB135-49), and Nature Science Foundation of Fujian Province, China (No. 2017J01125 and No. 2018J01106).

References

[1] J. Ahn and S. Kwak. Learning Pixel-level Semantic Affinity with Image-level Supervision for Weakly Supervised Semantic Segmentation. In *CVPR*, 2018.

[2] K. Alex, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[3] J. Amores. Multiple instance classification: Review, taxonomy and comparative study. *AI*, 2013.

[4] P. Arbeláez, J. Pont-Tuset, J. Barron, F. Marques, and J. Malik. Multiscale Combinatorial Grouping. In *CVPR*, 2014.

[5] L. Bazzani, A. Bergamo, D. Anguelov, and L. Torresani. Self-Taught Object Localization with Deep Networks. In *WACV*, 2016.

[6] A. Bearman, O. Russakovsky, V. Ferrari, and L. Fei-Fei. What's the Point: Semantic Segmentation with Point Supervision. In *ECCV*, 2016.

[7] A. J. Bency, H. Kwon, H. Lee, S. Karthikeyan, and B. S. Manjunath. Weakly Supervised Localization using Deep Feature Maps. In *ECCV*, 2016.

[8] H. Bilen, M. Pedersoli, and T. Tuytelaars. Weakly Supervised Object Detection with Posterior Regularization. In *BMVC*, 2014.

[9] H. Bilen, M. Pedersoli, and T. Tuytelaars. Weakly supervised object detection with convex clustering. In *CVPR*, 2015.

[10] H. Bilen and A. Vedaldi. Weakly Supervised Deep Detection Networks. In *CVPR*, 2016.

[11] A. Chaudhry, P. K. Dokania, and P. H. S. Torr. Discovering Class-Specific Pixels for Weakly-Supervised Semantic Segmentation. In *BMVC*, 2017.

[12] L.-C. Chen, A. Hermans, G. Papandreou, F. Schroff, P. Wang, and H. Adam. MaskLab: Instance Segmentation by Refining Object Detection with Semantic and Direction Features. In *CVPR*, 2018.

[13] 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.

[14] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *TPAMI*, 2017.

[15] R. G. Cinbis, J. Verbeek, and C. Schmid. Multi-fold MIL Training for Weakly Supervised Object Localization. In *CVPR*, 2014.

[16] R. G. Cinbis, J. Verbeek, and C. Schmid. Weakly Supervised Object Localization with Multi-fold Multiple Instance Learning. *TPAMI*, 2015.

[17] 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.

[18] T. Deselaers, B. Alexe, and V. Ferrari. Weakly supervised localization and learning with generic knowledge. *IJCV*, 2012.

[19] A. Diba, V. Sharma, A. Pazandeh, H. Pirsiavash, and L. Van Gool. Weakly Supervised Cascaded Convolutional Networks. In *CVPR*, 2017.

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

[21] R. Fan, Q. Hou, M.-M. Cheng, G. Yu, R. R. Martin, and S.-M. Hu. Associating Inter-Image Salient Instances for Weakly Supervised Semantic Segmentation. In *ECCV*, 2018.

[22] M. Gao, A. Li, R. Yu, V. I. Morariu, and L. S. Davis. C-WSL: Count-guided Weakly Supervised Localization. In *ECCV*, 2018.

[23] C. Ge and J. Wang. Fewer is More : Image Segmentation Based Weakly Supervised Object Detection with Partial Aggregation. In *BMVC*, 2018.

[24] W. Ge, S. Yang, and Y. Yu. Multi-Evidence Filtering and Fusion for Multi-Label Classification, Object Detection and Semantic Segmentation Based on Weakly Supervised Learning. In *CVPR*, 2018.

[25] S. Gidaris and N. Komodakis. Object detection via a multi-region & semantic segmentation-aware CNN model. In *ICCV*, 2015.

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

[27] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *AISTATS*, 2010.

[28] A. Gudi, N. van Rosmalen, M. Loog, and J. van Gemert. Object-Extent Pooling for Weakly Supervised Single-Shot Localization. In *BMVC*, 2017.

[29] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik. Simultaneous Detection and Segmentation. In *ECCV*, 2014.

[30] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask R-CNN. In *ICCV*, 2017.

[31] K. He, X. Zhang, S. Ren, and J. Sun. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. In *ECCV*, 2014.

[32] S. Hong, D. Yeo, S. Kwak, H. Lee, and B. Han. Weakly Supervised Semantic Segmentation using Web-Crawled Videos. In *CVPR*, 2017.

[33] G. Huang, Z. Liu, and K. Q. Weinberger. Densely Connected Convolutional Networks. In *CVPR*, 2017.

[34] Z. Huang, X. Wang, J. Wang, W. Liu, and J. Wang. Weakly-Supervised Semantic Segmentation Network with Deep Seeded Region Growing. In *CVPR*, 2018.

[35] N. Inoue, R. Furuta, T. Yamasaki, and K. Aizawa. Cross-Domain Weakly-Supervised Object Detection through Progressive Domain Adaptation. In *CVPR*, 2018.

[36] Z. Jie, Y. Wei, X. Jin, J. Feng, and W. Liu. Deep Self-Taught Learning for Weakly Supervised Object Localization. In *CVPR*, 2017.

[37] Kaiming He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *CVPR*, 2016.

[38] V. Kantorov, M. Oquab, M. Cho, and I. Laptev. Context-LocNet: Context-Aware Deep Network Models for Weakly Supervised Localization. In *ECCV*, 2016.

[39] A. Kolesnikov and C. H. Lampert. Seed, Expand and Constrain: Three Principles for Weakly-Supervised Image Segmentation. In *ECCV*, 2016.

[40] D. Li, J.-B. Huang, Y. Li, S. Wang, and M.-H. Yang. Weakly Supervised Object Localization with Progressive Domain Adaptation. In *CVPR*, 2016.

[41] Y. Li, L. Liu, C. Shen, and A. van den Hengel. Image Co-localization by Mimicking a Good Detector's Confidence Score Distribution. In *ECCV*, 2016.

[42] D. Lin, J. Dai, J. Jia, K. He, and J. Sun. ScribbleSup: Scribble-Supervised Convolutional Networks for Semantic Segmentation. In *CVPR*, 2016.

[43] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[44] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia. Path Aggregation Network for Instance Segmentation. In *CVPR*, 2018.

[45] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. SSD: Single Shot MultiBox Detector. In *ECCV*, 2016.

[46] J. Long, E. Shelhamer, and T. Darrell. Fully Convolutional Networks for Semantic Segmentation. In *CVPR*, 2015.

[47] D. Mahajan, R. Girshick, V. Ramanathan, K. He, M. Paluri, Y. Li, A. Bharambe, and L. van der Maaten. Exploring the Limits of Weakly Supervised Pretraining. In *ECCV*, 2018.

[48] D. P. Papadopoulos, J. R. R. Uijlings, F. Keller, and V. Ferrari. We don't need no bounding-boxes: Training object class detectors using only human verification. In *CVPR*, 2016.

[49] G. Papandreou, L.-C. Chen, K. Murphy, and A. L. Yuille. Weakly- and Semi-Supervised Learning of a DCNN for Semantic Image Segmentation. In *ICCV*, 2015.

[50] D. Pathak, P. Krähenbühl, and T. Darrell. Constrained Convolutional Neural Networks for Weakly Supervised Segmentation. In *ICCV*, 2015.

[51] P. O. Pinheiro and R. Collobert. From Image-level to Pixel-level Labeling with Convolutional Networks. In *CVPR*, 2015.

[52] X. Qi, Z. Liu, J. Shi, H. Zhao, and J. Jia. Augmented feedback in semantic segmentation under image level supervision. In *ECCV*, 2016.

[53] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You Only Look Once: Unified, Real-Time Object Detection. In *CVPR*, 2016.

[54] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *NIPS*, 2015.

[55] F. Saleh, A. Akbarian, Mohammad, Sadegh, M. Salzmann, L. Petersson, S. Gould, and J. M. Alvarez. Built-in Foreground/Background Prior for Weakly-Supervised Semantic Segmentation. In *ECCV*, 2016.

[56] T. Shen, G. Lin, C. Shen, and I. Reid. Bootstrapping the Performance of Webly Supervised Semantic Segmentation. In *CVPR*, 2018.

[57] Y. Shen, R. Ji, C. Wang, X. Li, and X. Li. Weakly Supervised Object Detection via Object-Specific Pixel Gradient. *TNNLS*, 2018.

[58] Y. Shen, R. Ji, S. Zhang, W. Zuo, and Y. Wang. Generative Adversarial Learning Towards Fast Weakly Supervised Detection. In *CVPR*, 2018.

[59] M. Shi, H. Caesar, and V. Ferrari. Weakly Supervised Object Localization Using Things and Stuff Transfer. In *ICCV*, 2017.

[60] M. Shi and V. Ferrari. Weakly Supervised Object Localization Using Size Estimates. In *ECCV*, 2016.

[61] W. Shimoda and K. Y. B. Distinct Class-Specific Saliency Maps for Weakly Supervised Semantic Segmentation. In *ECCV*, 2016.

[62] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. In *ICLR*, 2014.

[63] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *ICLR*, 2015.

[64] K. K. Singh, F. Xiao, and Y. J. Lee. Track and Transfer: Watching Videos to Simulate Strong Human Supervision for Weakly-Supervised Object Detection. In *CVPR*, 2016.

[65] H. O. Song, Y. J. Lee, S. Jegelka, and T. Darrell. Weakly-supervised Discovery of Visual Pattern Configurations. In *NIPS*, 2014.

[66] M. Tang, A. Djelouah, F. Perazzi, Y. Boykov, and C. Schroers. Normalized Cut Loss for Weakly-supervised CNN Segmentation. In *CVPR*, 2018.

[67] M. Tang, F. Perazzi, A. Djelouah, I. B. Ayed, C. Schroers, and Y. Boykov. On Regularized Losses for Weakly-supervised CNN Segmentation. In *ECCV*, 2018.

[68] P. Tang, X. Wang, S. Bai, W. Shen, X. Bai, W. Liu, and A. Yuille. PCL: Proposal Cluster Learning for Weakly Supervised Object Detection. *TPAMI*, 2018.

[69] P. Tang, X. Wang, X. Bai, and W. Liu. Multiple Instance Detection Network with Online Instance Classifier Refinement. In *CVPR*, 2017.

[70] P. Tang, X. Wang, A. Wang, Y. Yan, W. Liu, J. Huang, and A. Yuille. Weakly Supervised Region Proposal Network and Object Detection. In *ECCV*, 2018.

[71] P. Vernaza and M. Chandraker. Learning Random-Walk Label Propagation for Weakly-Supervised Semantic Segmentation. In *CVPR*, 2017.

[72] F. Wan, P. Wei, J. Jiao, Z. Han, and Q. Ye. Min-Entropy Latent Model for Weakly Supervised Object Detection. In *CVPR*, 2018.

[73] C. Wang, W. Ren, K. Huang, and T. Tan. Weakly Supervised Object Localization with Latent Category Learning. In *ECCV*, 2014.

[74] X. Wang, S. You, X. Li, and H. Ma. Weakly-Supervised Semantic Segmentation by Iteratively Mining Common Object Features. In *CVPR*, 2018.

[75] X. Wang, Z. Zhu, C. Yao, and X. Bai. Relaxed Multiple-Instance SVM with Application to Object Discovery. In *ICCV*, 2015.

[76] Y. Wei, X. Liang, Y. Chen, X. Shen, M.-M. Cheng, J. Feng, Y. Zhao, and S. Yan. STC: A Simple to Complex Framework for Weakly-supervised Semantic Segmentation. *TPAMI*, 2017.

[77] Y. Wei, Z. Shen, B. Cheng, H. Shi, J. Xiong, J. Feng, and T. Huang. TS2C: Tight Box Mining with Surrounding Segmentation Context for Weakly Supervised Object Detection. In *ECCV*, 2018.

[78] Y. Wei, H. Xiao, H. Shi, Z. Jie, J. Feng, and T. S. Huang. Revisiting Dilated Convolution: A Simple Approach for

Weakly- and Semi- Supervised Semantic Segmentation. In *CVPR*, 2018.

[79] J. Xu, A. G. Schwing, and R. Urtasun. Learning to segment under various forms of weak supervision. In *CVPR*, 2015.

[80] Z. Yan, J. Liang, W. Pan, J. Li, and C. Zhang. Weakly- and Semi-Supervised Object Detection with Expectation-Maximization Algorithm. *ArXiv*, 2017.

[81] H. Zhang, K. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, and A. Agrawal. Context Encoding for Semantic Segmentation. In *CVPR*, 2018.

[82] X. Zhang, J. Feng, H. Xiong, and Q. Tian. Zigzag Learning for Weakly Supervised Object Detection. In *CVPR*, 2018.

[83] X. Zhang, Y. Wei, J. Feng, Y. Yang, and T. Huang. Adversarial Complementary Learning for Weakly Supervised Object Localization. In *CVPR*, 2018.

[84] Y. Zhang, Y. Li, and B. Ghanem. W2F : A Weakly-Supervised to Fully-Supervised Framework for Object Detection. In *CVPR*, 2018.

[85] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning Deep Features for Discriminative Localization. In *CVPR*, 2016.

[86] Y. Zhou, Y. Zhu, Q. Ye, Q. Qiu, and J. Jiao. Weakly Supervised Instance Segmentation using Class Peak Response. In *CVPR*, 2018.

[87] Y. Zhu, Y. Zhou, Q. Ye, Q. Qiu, and J. Jiao. Soft Proposal Networks for Weakly Supervised Object Localization. In *ICCV*, 2017.