Parallel Detection-and-Segmentation Learning for Weakly Supervised Instance Segmentation

Yunhang Shen[1,2], Liujuan Cao[1]*, Zhiwei Chen[1], Baochang Zhang[3]
Chi Su[4], Yongjian Wu[2], Feiyue Huang[2], Rongrong Ji[1,5,6]
[1]Media Analytics and Computing Lab, Department of Artificial Intelligence,
School of Informatics, Xiamen University, 361005, China, [2]Tencent Youtu Lab, Shanghai, China
[3]Institute of Artificial Intelligence, Beihang University, Beijing, China
[4]KingSoft Cloud Co. Ltd., Beijing, China, [5]Institute of Artificial Intelligence,
Xiamen University, Xiamen, China, [6]Peng Cheng Laborotory, Shenzhen, China
{odysseyshen, littlekenwu, garyhuang}@tencent.com, {caoliujuan, rrji}@xmu.edu.cn
zhiweichen.xmu@gmail.com, bczhang@buaa.edu.cn, suchi@kingsoft.com

Abstract

Weakly supervised instance segmentation (WSIS) with only image-level labels has recently drawn much attention. To date, bottom-up WSIS methods refine discriminative cues from classifiers with sophisticated multi-stage training procedures, which also suffer from inconsistent object boundaries. And top-down WSIS methods are formulated as cascade detection-to-segmentation pipeline, in which the quality of segmentation learning heavily depends on pseudo masks generated from detectors. In this paper, we propose a unified parallel detection-and-segmentation learning (PDSL) framework to learn instance segmentation with only image-level labels, which draws inspiration from both top-down and bottom-up instance segmentation approaches. The detection module is the same as the typical design of any weakly supervised object detection, while the segmentation module leverages self-supervised learning to model class-agnostic foreground extraction, following by self-training to refine class-specific segmentation. We further design instance-activation correlation module to improve the coherence between detection and segmentation branches. Extensive experiments verify that the proposed method outperforms baselines and achieves the state-of-the-art results on PASCAL VOC and MS COCO.

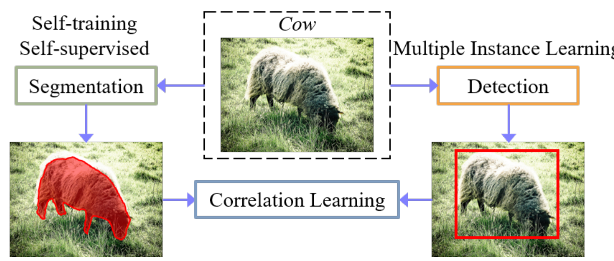

Figure 1: The overall flowchart of PDSL framework.

1. Introduction

Instance segmentation [1, 2] is one of the fundamental tasks in computer vision, which aims to simultaneously localize bounding boxes, classify target categories and estimate segmentation masks of object instances in images. Despite its significant progress in recent years, the dominant paradigms require a large number of training images with instance-level pixel-wise human annotations. However, collecting such fully-labelled training data is labor-intensive [3] and restricts applicability of instance segmentation in many downstream high-level vision tasks, ranging from autonomous driving, pose estimation to image synthesis. Thus, it has motivated the exploration of weakly supervised instance segmentation (WSIS), especially the setting where only image-level labels are used during training.

WSIS is an extremely challenging task with only a few attempts in previous literature. The bottom-up WSIS methods [4, 5, 6, 7, 8, 9] used classification networks to identify object instances from discriminative localization cues in the images. While this is a promising line of works, the initial localization cues are quite coarse and not consistent with object boundaries [10], which over-concentrates on discriminative parts of objects and under-estimates small instances [11]. Moreover, those methods suffer from sophisticated and multi-stage training processes to refine in-

*Corresponding author.

termediate results, which also rely on class activation maps and segmentation proposals. On the other hand, the top-down methods [10, 12] requires weakly supervise object detection (WSOD) to generate pseudo-ground-truth masks by leveraging gradient of image pixels w.r.t. detection results. Such detection-to-segmentation multi-task cascade causes the quality of pseudo masks heavily depending on WSOD, which limits further improvement with large margins. Although obtaining end-to-end pipeline, the top-down approaches have significantly inferior performance compared to the bottom-up ones with sophisticated training procedures prevailing in public benchmarks [13, 14].

To conquer the aforementioned limitations, we propose a unified parallel detection-and-segmentation learning (PDSL) framework to learn instance segmentation with only image-level labels, which draws inspiration from both top-down and bottom-up instance segmentation approaches. Our motivation is to parallel top-down detection and bottom-up segmentation via correlation learning in an end-to-end manner. The proposed PDSL framework has three advantages. First, compared to top-down WSIS methods, PDSL decouples the generation of pseudo-ground-truth masks from detectors and explores bottom-up object cues to learn segmentation, which models class-agnostic to class-specific foreground masks progressively. Second, compared to bottom-up methods, PDSL imposes bounding-box constraint from detection module on segmentation learning, which are encouraged to match up with object boundaries. Third, PDSL further collaborates detection module with segmentation learning by explicitly modelling correlations between them.

To this end, the proposed PDSL consists of three key components: object detection, image segmentation and correlation learning modules, as illustrated in Fig. 1. First, the detection branch is the same as the typical design of any top-down detectors learned from image-level labels. Second, the segmentation branch leverages self-supervised learning to model class-agnostic foreground extraction, which is followed by self-training to learn class-specific object segmentation constrained by bounding boxes. Third, to improve the coherence between detection and segmentation, we further propose instance-activation correlation learning, which impose a high correlation between two branches for activation of the same object instances. Extensive experiments on PASCAL VOC [13] and MS COCO [14] show that the proposed PDSL outperforms baseline models and achieves the state-of-the-art results. For the first time, we show that a top-down approach delivers competitive WSIS results.

The contributions of this work are three folds:

- We propose a cooperative parallel detection-and-segmentation learning framework to learn instance segmentation with only image-level labels. It introduces bottom-up object cues to top-down pipeline and disentangles segmentation supervision from detectors.
- The segmentation branch cooperates self-supervised learning and self-training to model from class-agnostic foreground extraction to class-specific object segmentation progressively, while the detection branch utilizes off-the-shelf WSOD methods to mine object in the form of bounding boxes.
- We further propose instance-activation correlation module to enhance the coherence between detection and segmentation branches.

2. Related Work

Weakly Supervised Instance Segmentation (WSIS). WSIS can be categorized into two groups according to training supervision, *i.e*., instance- and image-level labels. The first group mainly utilizes bounding-box annotations to supervise instance segmentation models. Khoreva *et al*. [15] applied box-driven segmentation techniques for bounding boxes individually to generate pixel-level labels and exploited recursive training as a de-noising strategy. Hsu *et al*. [16] introduced multiple instance learning to generate pseudo-ground-truth masks, which used tightness prior of bounding boxes to build positive and negative bags. Li *et al*. [17] extended [15] to iteratively refine pseudo masks with segmentation predictions during training. Arun *et al*. [18] proposed a joint probabilistic learning objective and conditional distributions of pseudo masks for different levels of weak supervision. Cholakkal *et al*. [19] constructed object category density maps with the spatial distribution of object-counting information to learn WSIS.

The second group further challenges WSIS problem with only image-level weak supervision. The early work commonly explored bottom-up methods, which contain sophisticated multi-stage training procedures. PRM [4] and IAM [5] utilized class response maps to extract discriminative localization cues via back-propagation, which leveraged segmentation proposals to generate instance masks. IRNet [6] and WISE [7] propagated coarse localization cues from CAM [20] to discover the entire object, which is further regarded as pseudo-ground-truth masks to train fully supervised models. Recent WSIS methods drifted from bottom-up to top-down manner [12, 21, 22], which detects and segments all object instances sequentially. Label-PEnet [12] developed multiple cascaded modules with curriculum learning strategy, which also relied on external models, *i.e*., Excitation BP [23], to compute segmentation masks at each stage. Kim *et al*. [21] proposed multi-task community learning to construct positive feedback loop and generated pseudo-ground-truth masks from CAM [20].

Unlike prevailing WSIS approaches, we propose a blender PDSL framework to learn top-down detection and

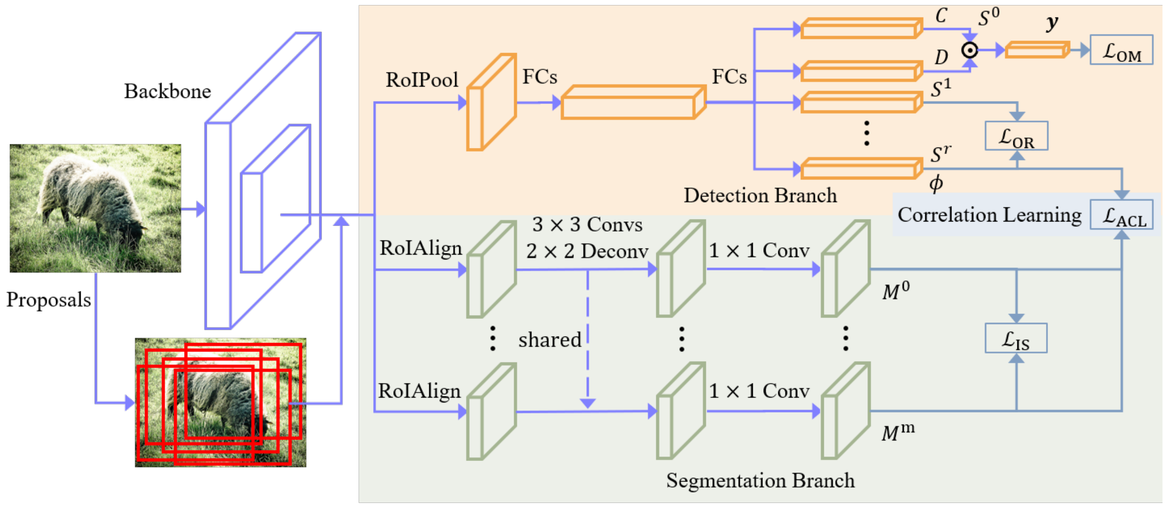

Figure 2: The figure illustrates the overall architecture of PDSL. The proposed PDSL consists of four components: backbone network, object detection, image segmentation and correlation learning modules.

bottom-up segmentation in parallel, which also captures activation-level interaction between detection and segmentation with correlation learning. The proposed PDSL gets rid of the sophisticated training procedures in the bottom-up approaches while achieving large performance improvement compared to previous top-down WSIS methods.

LIID [8] and S^4Net [9] utilized graph partition algorithms to assign pseudo-ground-truth labels for segmentation proposals, which are used to train fully supervised models. However, LIID [8] and S^4Net [9] required external instance-level segmentation models to compute salient instances as segmentation proposals, which contains additional ground-truth masks for training such segmentation.

Weakly Supervised Object Detection (WSOD). WSOD aims to predict object instance in the form of bounding boxes with weak supervision. WSDDN [24] selected proposals by parallel detection and classification branches in deep convolutional networks. Many work extended WSDDN [24] and leveraged contextual information [25, 10], attention mechanism [26] to suppress low-quality object proposals. Several different strategies to train the MIL model had been proposed in the literature [27, 28, 29, 30, 31, 32, 10]. Work in [33, 34, 35, 36, 37] treated the top-scoring proposals as supervision to train multiple instance refinement classifiers. Other different strategies [38, 39, 40, 41, 42, 43] are also proposed to generate pseudo-ground-truth bounding boxes and assign labels to proposals. The above framework is further improved by min-entropy prior [44, 45], gradient information [46, 47], continuation MIL [48], utilizing uncertainty [49, 50, 51], generative adversarial learning [52], spatial likelihood voting [53], objectness consistent [54, 55] and deep residual learning [56]. Methods in [57, 58, 59] trained object detection systems from different supervisions.

Collaboration mechanism between object detection and semantic segmentation is proposed to take advantages of the complementary interpretations of weakly supervised tasks [60, 61, 62, 54]. However, those approaches aimed to improve detection results with segmentation guidance from full-image masks, which are produced from off-the-shelf segmentation models [63, 6] or additional segmentation branches. Moreover, they neglected the correlation relationship between detection and segmentation, and only reported detection performance. As demonstrated in [21], a straightforward combination of such two techniques failed to achieve competitive results for instance segmentation.

3. The Proposed Method

3.1. Overall Framework

Given an image I and corresponding image-level labels $\mathbf{t} = [\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_{n^c}]$ during training, WSIS aims to estimate segmentation masks for all object instances. Here $\mathbf{t}$ is an binary vector, where $\mathbf{t}_c = 1$ denotes that image I contains the c^{th} target category, and otherwise, $\mathbf{t}_c = 0$. And n^c is the number of target categories. In this paper, we propose a unified parallel detection-and-segmentation learning (PDSL) framework to learn instance segmentation with only image-level labels, which draws inspiration from both top-down and bottom-up instance segmentation approaches. As illustrated in Fig 2, the proposed PDSL consists of four modules: backbone network, object detection, image segmentation and correlation learning modules. The backbone network first outputs full-image feature maps of input I. Then we use RoIPool [64] and RoIAlgin [65] layers to extract pooled features for pre-computed object proposals [66, 67], which are the inputs of object detection and image segmentation branches, respectively. Specifically, a top-down detection branch mines object instances by classifying and refining object proposals, while the image segmentation branch explores bottom-up object cues to learn class-agnostic and class-specific segmentation within bounding boxes via self-supervised learning and self-training. As both branches process the same proposals, we further introduce a correlation learning module to enhance the coherence between detection and segmentation. During training, we have following objective function

$$\mathcal{L} = \mathcal{L}_{\mathrm{OD}} + \mathcal{L}_{\mathrm{IS}} + \mathcal{L}_{\mathrm{CL}}, \tag{1}$$

where $\mathcal{L}_{\mathrm{OD}}$ and $\mathcal{L}_{\mathrm{IS}}$ are the loss functions of object detection and image segmentation, respectively. And $\mathcal{L}_{\mathrm{CL}}$ is the loss function of correlation learning.

3.2. Object Detection Branch

We follow the multiple-instance learning [68] pipeline in deep convolutional networks and utilize two-stream WSDDN [24] algorithms for object detection branch. Given pooled features from RoIPool [64] layer, we extract proposal features by two fully-connected layers, each of which is followed by ReLU activation and dropout layer. Then the proposal features are forked into two streams, *i.e.*, classification and detection stream, to produce two score matrices $C, D \in \mathbb{R}^{n^{\mathrm{p}} \times n^{\mathrm{c}}}$ by another two fully-connected layers, respectively, where n^{p} is the number of proposals. Both score matrices are normalized by softmax functions $\sigma(\cdot)$ over categories and proposals, respectively. Finally, the element-wise product of the output of the two streams is again a score matrix: $S^0 = \sigma(C) \odot \sigma(D^T)^T$. To acquire image-level multi-label classification scores, a sum pooling is applied: $\mathbf{y}_c = \sum_{r=1}^{n^{\mathrm{p}}} S^0_{rc}$. Then we employ multi-label cross-entropy loss function to utilize image-level labels as

$$\mathcal{L}_{\mathrm{OM}} = -\sum_{c=1}^{n^{\mathrm{c}}} \Big\{ \mathbf{t}_c \log \mathbf{y}_c + (1 - \mathbf{t}_c) \log(1 - \mathbf{y}_c) \Big\}. \tag{2}$$

To further reduce mis-localizations, we tweak the similar ideal from OICR [35] to refine detection results via multiple detection refinement heads. To this end, each head has proposal classification and bounding-box regression subnets, which enables to refine both bounding-box scores and coordinates. In details, it produces classification scores $S^{\mathrm{r}} \in \Re^{n^{\mathrm{p}} \times (n^{\mathrm{c}}+1)}$ and new bounding boxes $B^{\mathrm{r}} \in \Re^{n^{\mathrm{p}} \times n^{\mathrm{c}} \times 4}$ for the r^{th} refinement head, where $n^{\mathrm{c}} + 1$ indicates n^{c} object categories and 1 background category.

During training, for the r^{th} head and the c^{th} category that $\mathbf{t}_c = 1$, the highest-score bounding box from previous prediction is selected as pseudo-ground-truth boxes and assigns positive/negative labels for each proposal. Thus, the corresponding objective function is

$$\begin{aligned} \mathcal{L}^{\mathrm{r}}_{\mathrm{OR}} = \sum_{p=1}^{n^{\mathrm{p}}} & \mathbf{y}_{\mathbf{t}^{\mathrm{r}}_p} \mathcal{L}_{\mathrm{CE}}(S^{\mathrm{r}}_p, \mathbf{t}^{\mathrm{r}}_p) + \\ & [\mathbf{t}^{\mathrm{r}}_p \geq 1] \mathbf{y}_{\mathbf{t}^{\mathrm{r}}_p} \mathcal{L}_{\mathrm{smoothL1}}(B^{\mathrm{r}}_{p\mathbf{t}^{\mathrm{r}}_p}, \hat{B}^{\mathrm{r}}_p), \end{aligned} \tag{3}$$

where $\mathbf{t}^{\mathrm{r}} \in \Re^{n^{\mathrm{p}}}$ and $\hat{B}^{\mathrm{r}} \in \Re^{n^{\mathrm{p}} \times 4}$ are the classification and regression targets for object proposals in the r^{th} head, respectively. L_{CE} is the softmax cross-entropy loss, and $\mathcal{L}_{\mathrm{smoothL1}}$ is the smooth L1 loss [64]. The iverson bracket indicator function $[\mathbf{t}^{\mathrm{r}}_p \geq 1]$ evaluates to 1 when $\mathbf{t}^{\mathrm{r}}_p \geq 1$ and 0 otherwise. With above definition, the overall objective function for object detection module is

$$\mathcal{L}_{\mathrm{OD}} = \mathcal{L}_{\mathrm{OM}} + \sum_{r=1}^{n^{\mathrm{r}}} \mathcal{L}^{\mathrm{r}}_{\mathrm{OR}}, \tag{4}$$

where n^{r} is the number of detection refinement heads. During testing, the average output of all heads is used.

3.3. Image Segmentation Branch

Image segmentation branch aims to predict instances masks given bounding boxes in images. In fully supervised learning, image segmentation can directly learn ground-truth masks within positive bounding boxes [65]. However, recent top-down WSIS methods, *i.e.*, WSJDS [10] and Label-PEnet [12], back-propagated detection results to images to generate object heatmaps, which are then post-processed as supervision for segmentation learning. Thus, the quality of pseudo-ground-truth masks is strongly tied to the performance of object detection, which limits further improvement with large margins.

In this paper, we formulate image segmentation as foreground extraction via a progressive class-agnostic to class-specific strategy. Particularly, we first leverage self-supervised learning to learn class-agnostic foreground segmentation from bottom-up object cues within bounding boxes, which disentangles the pseudo-ground-truth mask generation from detection module. Then, the class-agnostic mask predictions are treated as supervision to learn class-specific segmentation branches via self-training. To this end, image segmentation module consists of multiple mask prediction heads with the same structure. In detail, image segmentation module has 4 convolutional layers with 3×3 kernels and 256 channels to extract feature maps, which followed by a deconvolutional layer with 2×2 kernels and n^{m} final prediction layer with 1×1 kernels. Given a set of object proposals, the objective function $\mathcal{L}_{\text{IS}}$ is defined as

$$\begin{aligned}\mathcal{L}_{\text{IS}} = &\sum_{p=1}^{n^{\text{p}}} \mathcal{L}_{\text{BCE}}(M_p^0, \hat{M}_p^0) + \\ &\sum_{p=1}^{n^{\text{p}}} \sum_{c=1}^{n^{\text{c}}} [\mathbf{t}_p^{n^{\text{r}}} = c] \mathbf{y}_{\mathbf{t}_p^{n^{\text{r}}}} \mathcal{L}_{\text{BCE}}(M_{pc}^1, M_p^0) + \\ &\sum_{m=2}^{n^{\text{m}}} \sum_{p=1}^{n^{\text{p}}} \sum_{c=1}^{n^{\text{c}}} [\mathbf{t}_p^{n^{\text{r}}} = c] \mathbf{y}_{\mathbf{t}_p^{n^{\text{r}}}} \mathcal{L}_{\text{BCE}}(M_{pc}^{\text{m}}, M_{pc}^{\text{m}-1}),\end{aligned} \tag{5}$$

where the first term and the last two terms are loss function of class-agnostic and multiple class-specific mask heads, respectively. And $\mathcal{L}_{\text{BCE}}(M, \hat{M}) = -\sum_{i,j} \hat{M}_{ij} \log M_{ij} - (1 - \hat{M}_{ij}) \log(1 - M_{ij})$ is the binary cross-entropy loss. Specifically, M_p^0 and $\hat{M}_p^0$ are the mask prediction and pseudo-ground-truth targets for the p^{th} proposal in class-agnostic mask head. And M_{pc}^{m} denotes the prediction masks for the p^{th} proposal and the c^{th} category in class-specific mask heads. Different to class-agnostic mask head, class-specific mask heads only compute losses for the categories existed in the images, which are weighted by the image-level classification scores $\mathbf{y}_{\mathbf{t}_p^{n^{\text{r}}}}$.

To acquire initial pseudo-ground-truth masks $\hat{M}^0$ for class-agnostic mask head, we use unsupervised GrabCut [69] methods to extract bottom-up object cues as foreground segmentation given bounding boxes. We are not restricted with the algorithms that generate object cues from input images. However, extract foreground segmentation for all proposals is computationally inefficient during training. As a large number of proposals are necessary to achieve a reasonable recall rate and good performance. Recall that object proposals are always redundant and highly overlap each other, which makes their masks shareable. Thus, we introduce a seed sample acquisition strategy. Specifically, we first pick the highest-score object proposal for each category that appears in the category-label. We then sort the object proposals according to IoU overlaps with the selected proposal. After that, the top n^{seed} proposals are sampled as seeds to estimate foreground segmentation for pseudo-ground-truth masks, where $n^{\text{seed}} \ll n^{\text{p}}$. Finally, the pseudo-ground-truth masks of the rest proposals are the same as that of seed proposals with the highest box IoU overlap. Despite its simplicity, our experiment shows that, even with a small $n^{\text{seed}} = 10$ for each category, the generated pseudo-ground-truth masks still enable class-agnostic mask head to learn high-quality foreground segmentation.

3.4. Correlation Learning Module

Theoretically, the loss functions of detection and segmentation task lead to complementary knowledge [10]. MIL-based WSOD explicitly penalizes all false positives, and counts a prediction as correct as long as it has IoU with ground truth over a threshold. This brings clean background with few false positives, but also lacks sensitivity to fine tune the localizer. On the other hand, for segmentation, the lack of explicit penalty on false positives often results in noisy background. But the fine granularity gives it better precision on ambiguous regions to guide the object localizer. So these two tasks complement well with each other. Our motivation of correlation learning is to exploit complementary knowledge learned from individual tasks by enhancing the coherence between detection and segmentation. Our ablation study also demonstrates that correlation learning is vital to achieve high performance for parallel detection-and-segmentation learning.

Although PDSL learns detection and segmentation in parallel, both modules are applied to feature maps cropped by the same proposals from fully-image feature maps. Thus, we design an instance-activation correlation learning with overall objective function formulated as

$$\mathcal{L}_{\text{CL}} = \sum_{m=1}^{n^{\text{m}}} \mathcal{L}_{\text{ACL}}^{\text{m}}, \tag{6}$$

where $\mathcal{L}_{\text{ACL}}^{\text{m}}$ is the loss functions of instance-activation correlations for the m^{th} segmentation branch.

The instance-activation correlation learning requires consistent prediction activation between detection and segmentation for the same proposals. We first compute the proposal activation a_p^{m} of the p^{th} proposals and the m^{th} mask head using Log-Sum-Exp [70] as

$$a_p^{\text{m}} = \frac{1}{\tau} \log\big(\frac{1}{hw} \sum \exp(\tau M_{p\mathbf{t}_p^{n^{\text{r}}}}^{\text{m}})\big), \tag{7}$$

where h, w are the spatial size of prediction mask M_{pc}^{r}, and hyper-parameter $\tau = 5$ controls the smooth. The instance activation from detection module is the predicted scores from the last object refinement heads: $\hat{a}_p = S_{p\mathbf{t}_p^{n^{\text{r}}}}^{n^{\text{r}}}$. Thus, the instance-activation correlation loss $\mathcal{L}_{\text{ACL}}^{\text{m}}$ for the m^{th} mask refinement head computes p-norm distance as

$$\mathcal{L}_{\text{ACL}}^{\text{m}} = \frac{1}{n^{\text{p}}} \sum_{p=1}^{n^{\text{p}}} ||a_p^{\text{m}} - \hat{a}_p||_2. \tag{8}$$

4. Quantitative Evaluations

4.1. Datasets

We evaluate the proposed method on PASCAL VOC 2012 [13] and MS COCO [14]. PASCAL VOC 2012 consists of 20 target categories as well as a background category. We follow [4, 6, 8] to utilize the main *trainval* subset, excluding segmentation *val* images, to train our models. We evaluate our approach and baseline models using the *1,449*

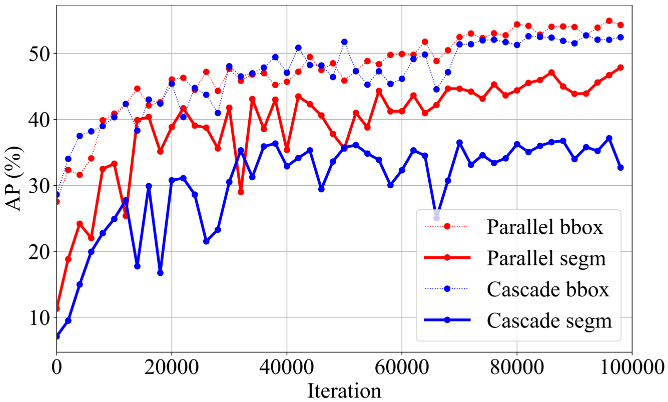

Figure 3: Object detection and instance segmentation performance on VOC 2012 for parallel and cascade learning.

segmentation val images. MS COCO dataset consists of *80* target categories. We follow [9] to train on the standard *train* set and evaluate on the *val* set. Note that only image-level labels are used for model training.

Table 1: Ablation study of PDSL on PASCAL VOC 2012 instance segmentation.

n^{r}	n^{m}	n^{seed}	$\mathcal{L}_{\mathrm{ACL}}$	mAP	$m\mathrm{AP}_{0.50}$	$m\mathrm{AP}_{0.75}$	$m\mathrm{AP}_{\mathrm{S}}$	$m\mathrm{AP}_{\mathrm{M}}$	$m\mathrm{AP}_{\mathrm{L}}$
0	4	10	✓	8.7	31.8	2.5	0.6	2.8	13.6
1	4	10	✓	14.3	40.6	8.7	1.8	6.0	19.1
2	4	10	✓	19.1	45.4	14.7	2.4	11.1	26.1
3	4	10	✓	20.2	47.3	15.5	2.5	12.9	28.3
4	4	10	✓	20.8	47.8	15.6	2.3	12.7	28.7
5	4	10	✓	20.9	47.8	15.6	2.2	12.3	28.8
4	0	10	✓	16.0	45.4	11.7	1.4	10.1	18.1
4	1	10	✓	16.6	45.8	12.1	1.1	10.9	20.1
4	2	10	✓	16.6	46.4	12.2	1.4	11.3	20.1
4	3	10	✓	17.3	46.9	13.0	1.8	12.9	22.8
4	4	10	✓	20.8	47.8	15.6	2.3	12.7	28.7
4	5	10	✓	20.8	47.9	15.0	2.1	13.1	28.9
4	4	1	✓	15.1	38.5	9.3	1.2	7.8	22.1
4	4	10	✓	20.8	47.8	15.6	2.3	12.7	28.7
4	4	50	✓	21.0	47.8	15.6	2.6	12.9	28.4
4	4	100	✓	20.9	48.0	15.0	2.3	12.7	28.5
4	4	10		17.8	43.7	11.6	1.6	9.1	26.1
4	4	10	✓	20.8	47.8	15.6	2.3	12.7	28.7

4.2. Evaluation Protocol

For the evaluation metrics of instance segmentation, we report the standard COCO metrics [14], which is mean average precision (AP) over IoU thresholds. For object detection on Pascal VOC, we follow the standard PASCAL VOC protocol to report the *m*AP at 50% Intersection-over-Union (IoU) of the detected boxes with the ground-truth ones. We also report CorLoc to indicate the percentage of images in which a method correctly localizes an object of the target category according to the PASCAL criterion.

4.3. Implementation Details

We implement our method using the PyTorch framework. All backbones are initialized with the weights pre-trained on ImageNet ILSVRC [71]. We use synchronized SGD training on 4 GPUs. A mini-batch involves 1 images per GPU. We use a learning rate of 0.01, momentum of 0.9, and dropout rate of 0.5. We use a step learning rate decay schema with decay weight of 0.1 and step size of $70,000$ iterations. The total number of training iterations is $100,000$. We adopt $240,000$ training iterations for MS COCO. In the multi-scale setting, we use scales range from 480 to 1216 with stride 32. 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. A random crop with 0.9 of the size of the original images is applied. We use MCG [67] to generate object proposals for all experiments, including our implementation of baseline methods. We set the maximum number of proposals in an image to be $2,000$. The test scores are the average of scales of $\{480, 576, 672, 768, 864, 960, 1056, 1152\}$ and flips. Detection results are post-processed by NMS with threshold of 0.5. We use the following parameter settings in all the experiments, unless specified otherwise. We set the both hyper-parameters n^{r} and n^{m} in Eq. 4 and 5 to 4. For the seed sample acquisition strategy, we set the number of sampled proposals n^{seed} to 10.

4.4. Ablation Study

Before the comparison with other competitors, we perform several ablation studies to evaluate the effectiveness of different design choices and parameter settings. All ablation studies are conducted on the PASCAL VOC 2012 instance segmentation. Here, we use ResNet18-WS [56] as backbone to save time if not mentioned. When tuning each hyper-parameter, other parameters are kept as default.

Parallel *vs*. cascade learning strategies. Recall that previous top-down methods [10, 12] are based on multi-task cascade and utilized the gradient of detection results with respect to images to generate pseudo masks. The proposed PDSL method leverages correlation learning and self-supervised learning with bottom-up cues to model segmentation in parallel with detection. Thus, we also show the influence of different learning strategies, *i.e*., parallel and cascade learning, as plotted in Fig. 3. We compute the AP results for object detection and instance segmentation for each $2,000$ iterations with single-scale testing. It is obvious that parallel learning provides superior performance for instance segmentation compared to cascade learning. As detection-to-segmentation multi-task cascade causes the quality of pseudo masks heavily relying on object detection.

The number n^{r} of detection refinement heads. The detection refinement heads output bounding boxes for segmentation during testing, which heavily influence the performance of instance segmentation. The hyperparameter n^{r} in Equ. 4 controls the number of refinement heads. Different settings of n^{r} are evaluated in Tab. 1. When we have $n^{\mathrm{r}} = 0$, the second term of loss function $\mathcal{L}_{\mathrm{OD}}$ in Equ. 4 are omitted. We can see that the results of this setting are worse than using more refinement heads, demonstrating that the

Table 2: Comparison with the state-of-the-art methods on PASCAL VOC 2012 instance segmentation. The terms $\mathcal{M}$, $\mathcal{B}$, $\mathcal{C}$ and $\mathcal{I}$ denote pixel-level, bounding-box-level, object-count and image-level labels, respectively.

Method	Supervision	Backbone	$m\mathrm{AP}_{0.25}$	$m\mathrm{AP}_{0.50}$	$m\mathrm{AP}_{0.70}$	$m\mathrm{AP}_{0.75}$
Mask R-CNN [65]	$\mathcal{M}$	ResNet-101	76.7	67.9	52.5	44.9
Khoreva *et al.* [15]	$\mathcal{B}$	VGG-16	-	44.8	-	16.3
Cholakkal *et al.* [19]	$\mathcal{C}$	ResNet-50	48.5	30.2	-	14.4
Hsu *et al.* [16]	$\mathcal{B}$	ResNet-101	75.0	58.9	30.4	21.6
Arun *et al.* [18]	$\mathcal{B}$	ResNet-101	73.8	58.2	34.3	32.1
Bottom-up multi-stage-training WSIS methods						
CAM MCG [20]	$\mathcal{I}$	VGG-16	20.4	7.8	-	2.5
MELM MGC [45]	$\mathcal{I}$	VGG-16	36.9	22.9	-	8.4
SPN MGC [72]	$\mathcal{I}$	VGG-16	26.4	12.7	-	4.4
PRM [4]	$\mathcal{I}$	ResNet-50	44.3	26.8	-	9.0
IAM [5]	$\mathcal{I}$	ResNet-50	45.9	28.8	-	11.9
IRNet [6]	$\mathcal{I}$	ResNet-50	-	46.7	23.5	-
WISE [7]	$\mathcal{I}$	ResNet-50	49.2	41.7	-	23.7
Arun *et al.* [18]	$\mathcal{I}$	ResNet-50	**59.7**	**50.9**	**30.2**	**28.5**
LIID [8]	$\mathcal{I}$	ResNet-50	–	48.4	–	24.9
Top-down end-to-end WSIS methods						
Label-PEnet [12]	$\mathcal{I}$	VGG-16	49.1	30.2	-	12.9
Kim *et al.* [21]	$\mathcal{I}$	VGG-16	52.4	28.9	-	5.2
Kim *et al.* [21]	$\mathcal{I}$	ResNet-50	57.0	35.7	–	5.8
JTSM [22]	$\mathcal{I}$	ResNet18-WS	–	44.2	–	12.0
PDSL	$\mathcal{I}$	ResNet18-WS	58.6	47.8	–	**15.6**
		ResNet50-WS	**59.3**	49.6	–	12.7
		ResNet101-WS	59.2	**49.7**	–	13.1

Table 3: Comparison with the state-of-the-art methods on MS COCO instance segmentation. The terms $\mathcal{M}$, $\mathcal{S}$ and $\mathcal{I}$ denote pixel-level, instance saliency and image-level labels, respectively.

Method	Supervision	Backbone	$m\mathrm{AP}$	$m\mathrm{AP}_{0.50}$	$m\mathrm{AP}_{0.75}$	$m\mathrm{AP}_S$	$m\mathrm{AP}_M$	$m\mathrm{AP}_L$
Mask R-CNN [65]	$\mathcal{M}$	ResNet101	35.7	58.0	37.8	15.5	38.1	52.4
Bottom-up multi-stage-training WSIS methods with external supervision								
Fan *et al.* [9]	$\mathcal{I},\mathcal{S}$	ResNet101	13.7	25.5	13.5	0.7	15.7	26.1
LIID *et al.* [8]	$\mathcal{I},\mathcal{S}$	ResNet50	**16.0**	**27.1**	**16.5**	**3.5**	**15.9**	**27.7**
Top-down end-to-end WSIS methods								
WS-JDS [10]	$\mathcal{I}$	VGG16	6.1	11.7	**5.5**	**1.5**	**7.1**	12.2
JTSM [22]	$\mathcal{I}$	ResNet18-WS	6.1	12.1	5.0	0.1	3.0	**12.6**
PDSL	$\mathcal{I}$	ResNet18-WS	**6.3**	**13.1**	5.0	0.1	3.6	12.2

detection refinement is very helpful for instance segmentation predictions. When $n^{\mathrm{r}} \geq 4$, the performance gains are margin. Therefore, we use 4 as the default values for n^{r}.

The number n^{m} of class-specific mask heads. To analyze how AP varies with the number of segmentation refinement, we range hyper-parameter n^{m} in Equ. 5 from 0 to 5 When we have $n^{\mathrm{m}} = 0$, the second and third terms of loss function $\mathcal{L}_{\mathrm{IS}}$ in Equ. 5 are omitted, as we only have class-agnostic mask head. In Tab. 1, the performances are improved consistently for all metrics with hyper-parameter n^{m} increasing. For $n^{\mathrm{m}} \geq 4$, the performance improvement is relatively small. Therefore, we set n^{m} to 4 by default.

The number n^{seed} of sampled proposals for seed sample acquisition strategy. We continue by evaluating the effect of hyper-parameter n^{seed}. Recall that we only sample n^{seed} proposals to estimate foreground segmentation using unsupervised traditional methods. And generating such pseudo-ground-truth masks is time-consuming during training. Thus, we seek to balance the quality of pseudo-ground-truth masks and training speed by tuning hyper-parameter n^{seed}. As shown in Tab. 1, when only the highest-score proposal is used, *i.e.*, $n^{\mathrm{seed}} = 1$, the quality of learned masks drop dramatically. We can observe that compared with $n^{\mathrm{seed}} = 1$, even just using only 10 object proposals boosts the performance a lot, which confirms the effectiveness of leveraging object cues for segmentation module. We also find that more seed proposals help to improve pseudo-ground-truth masks for class-agnostic mask head. We set the default values for n^{seed} as 10, which provides a good balance between final performance and training time.

The instance-activation correlation learning loss $\mathcal{L}_{\mathrm{ACL}}$. To understand the importance of the correlation

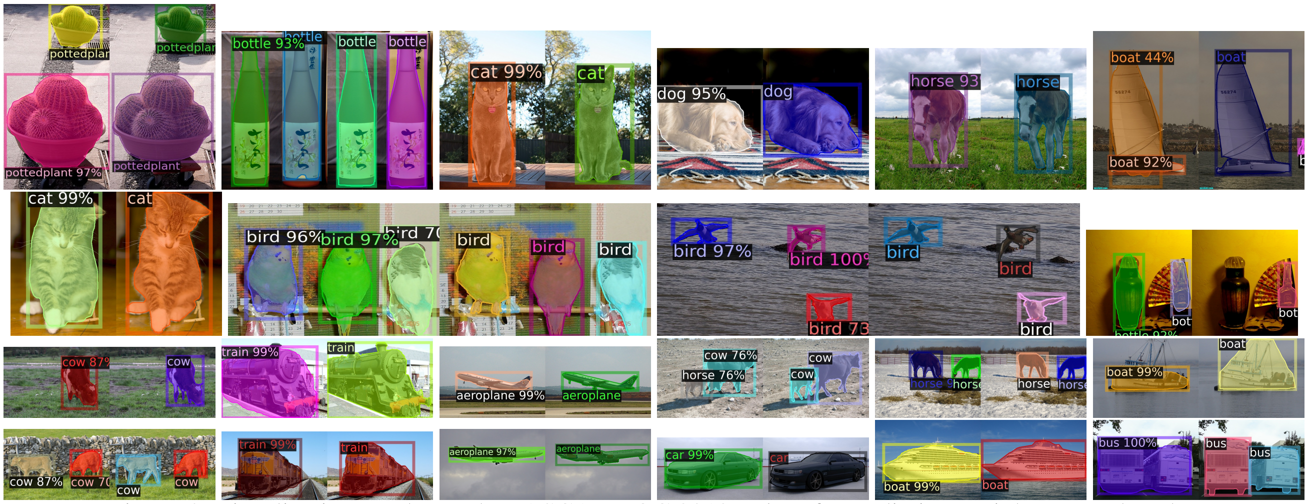

Figure 4: Visualization results on the PASCAL VOC 2012 *val*. Left: predictions, Right: ground truths.

learning, we evaluate the influence of different loss functions in this module. The instance-activation correlation loss $\mathcal{L}_{\mathrm{ACL}}$ punishes the activation diversity of the same proposals between detection and segmentation modules. As we can see in Tab. 1, results can be improved by instance-activation correlation learning.

4.5. Comparison with State of the Arts (SOTAs)

We comprehensively evaluate our method with three ResNet-WS [56] backbones in our experiments, which are variation of ResNet [73]. It should be noted that our default hyper-parameters are not the best setting according to Tab. 1. Comparisons with recent state-of-the-art methods on PASCAL VOC 2012 and MS COCO are listed in Tab. 2 and 3. Some previous methods achieve high performance, thanks to the specially designed inter-pixel relation module [6], graph partition algorithm [8, 9] and salient detector [9, 8]. Unlike previous methods [7, 6, 4, 5], our PDSL does not rely on require sophisticated and multiple sequential training process, *i.e.*, fully-supervised model retraining [7, 6], class activation map module [6, 7] and segmentation proposals [4, 5]. Instead, we utilize a powerful unified parallel detection-and-segmentation learning framework and correlation learning module to capture intra-modular and inter-modular dependencies across separate branches. Thus, we achieve consistent accuracy gain over existed methods and set the new state-of-the-art results.

The qualitative results on the PASCAL VOC 2012 *val* are shown in Fig. 4. As can be observed in the first five columns, our approach outputs semantically meaningful and precise predictions despite the existence of complex object appearances and challenging background contents. It demonstrated the effectiveness of the proposed unified parallel detection-and-segmentation learning framework. We further visualize our failure mode in the last column, mainly resulting from confusion with similar objects, localization error and failing to distinguish multiple instances.

Each iteration of PDSL with ResNet50 takes about 1 second for forward-backward propagation on GTX 1080Ti GPUs and several seconds on CPUs to extract bottom-up object cues. Thus, the total training times are about 12 days for PASCAL VOC. Noted that parallel GPU computing of bottom-up cues can further reduce training time. We can also use pre-computed GrabCut masks, segmentation proposals and attention maps as bottom-up object cues. During inference, PDSL runs at 539 ms per image including NMS.

5. Conclusion

In this paper, we propose a unified parallel detection-and-segmentation learning (PDSL) framework to learn instance segmentation with only image-level labels, which draws inspiration from both top-down and bottom-up instance segmentation approaches. In order to improve the coherence between detection and segmentation branches, we further propose instance-activation correlation learning, which impose a high correlation on activation between two branches for the same object instance. For the first time, we show a top-down WSIS approach could deliver the state-of-the-art results on both PASCAL VOC and MS COCO.

6. Acknowledgment

This work is supported by the National Science Fund for Distinguished Young Scholars (No.62025603), the National Natural Science Foundation of China (No.U1705262, No.62072386, No.62072387, No.62072389, No.62002305, No.61772443, No.61802324 and No.61702136), Guangdong Basic and Applied Basic Research Foundation (No.2019B1515120049), the Fundamental Research Funds for the central universities (No.20720200077, No.20720200090 and No.20720200091) and the Beijing Nova Program under Grant Z201100006820023.

References

[1] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Simultaneous Detection and Segmentation. In *European Conference on Computer Vision (ECCV)*, 2014. 1

[2] Pedro O. Pinheiro, Ronan Collobert, and Piotr Dollar. Learning to Segment Object Candidates. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2015. 1

[3] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes Dataset for Semantic Urban Scene Understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1

[4] Yanzhao Zhou, Yi Zhu, Qixiang Ye, Qiang Qiu, and Jianbin Jiao. Weakly Supervised Instance Segmentation using Class Peak Response. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 2, 5, 7, 8

[5] Yi Zhu, Yanzhao Zhou, Huijuan Xu, Qixiang Ye, David Doermann, and Jianbin Jiao. Learning Instance Activation Maps for Weakly Supervised Instance Segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2, 7, 8

[6] Jiwoon Ahn, Sunghyun Cho, and Suha Kwak. Weakly Supervised Learning of Instance Segmentation with Inter-pixel Relations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2, 3, 5, 7, 8

[7] Issam H. Laradji, David Vazquez, and Mark Schmidt. Where are the Masks: Instance Segmentation with Image-level Supervision. In *The British Machine Vision Conference (BMVC)*, 2019. 1, 2, 7, 8

[8] Yun Liu, Yu-huan Wu, Peisong Wen, Yujun Shi, Yu Qiu, and Ming-ming Cheng. Leveraging Instance- , Image- and Dataset-Level Information for Weakly Supervised Instance Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020. 1, 3, 5, 7, 8

[9] Ruochen Fan, Qibin Hou, Ming-Ming Cheng, Gang Yu, Ralph R. Martin, and Shi-Min Hu. Associating Inter-Image Salient Instances for Weakly Supervised Semantic Segmentation. In *European Conference on Computer Vision (ECCV)*, 2018. 1, 3, 6, 7, 8

[10] Yunhang Shen, Rongrong Ji, Yan Wang, Yongjian Wu, and Liujuan Cao. Cyclic Guidance for Weakly Supervised Joint Detection and Segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2, 3, 4, 5, 6, 7

[11] Nikita Araslanov and Stefan Roth. Single-Stage Semantic Segmentation from Image Labels. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1

[12] Weifeng Ge, Sheng Guo, Weilin Huang, and Matthew R. Scott. Label-PEnet: Sequential Label Propagation and Enhancement Networks for Weakly Supervised Instance Segmentation. In *IEEE International Conference on Computer Vision (ICCV)*, 2019. 2, 4, 6, 7

[13] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision (IJCV)*, 2010. 2, 5

[14] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision (ECCV)*, 2014. 2, 5, 6

[15] Anna Khoreva, Rodrigo Benenson, Jan Hosang, Matthias Hein, and Bernt Schiele. Simple Does It: Weakly Supervised Instance and Semantic Segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2, 7

[16] Cheng-chun Hsu, Yen-yu Lin, and Yung-yu Chuang. Weakly Supervised Instance Segmentation using the Bounding Box Tightness Prior. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2019. 2, 7

[17] Qizhu Li, Anurag Arnab, and Philip H. S. Torr. Weakly- and Semi-Supervised Panoptic Segmentation. In *European Conference on Computer Vision (ECCV)*, 2018. 2

[18] Aditya Arun, C. V. Jawahar, and M. Pawan Kumar. Weakly Supervised Instance Segmentation by Learning Annotation Consistent Instances. In *European Conference on Computer Vision (ECCV)*, 2020. 2, 7

[19] Hisham Cholakkal, Guolei Sun, Fahad Shahbaz Khan, and Ling Shao. Object counting and instance segmentation with image-level supervision. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 7

[20] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning Deep Features for Discriminative Localization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2, 7

[21] Seohyun Kim, Jaedong Hwang, Jeany Son, and Bohyung Han. Weakly Supervised Instance Segmentation by Deep Multi-Task Community Learning. *arXiv*, 2020. 2, 3, 7

[22] Yunhang Shen, Liujuan Cao, Zhiwei Chen, Feihong Lian, Baochang Zhang, Chi Su, Yongjian Wu, Feiyue Huang, and Rongrong Ji. Toward Joint Thing-and-Stuff Mining for Weakly Supervised Panoptic Segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 7

[23] Jianming Zhang, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down Neural Attention by Excitation Backprop. In *European Conference on Computer Vision (ECCV)*, 2016. 2

[24] Hakan Bilen and Andrea Vedaldi. Weakly Supervised Deep Detection Networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3, 4

[25] Vadim Kantorov, Maxime Oquab, Minsu Cho, and Ivan Laptev. ContextLocNet: Context-Aware Deep Network Models for Weakly Supervised Localization. In *European Conference on Computer Vision (ECCV)*, 2016. 3

[26] Eu Wern Teh and Yang Wang. Attention Networks for Weakly Supervised Object Localization. In *The British Machine Vision Conference (BMVC)*, 2016. 3

[27] Hyun Oh Song, Ross Girshick, Stefanie Jegelka, Julien Mairal, Zaid Harchaoui, and Trevor Darrell. On learning to localize objects with minimal supervision. In *International Conference on Machine Learning (ICML)*, 2014. 3

[28] Ramazan Gokberk Cinbis, Jakob Verbeek, and Cordelia Schmid. Multi-fold MIL Training for Weakly Supervised Object Localization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 3

[29] Chong Wang, Weiqiang Ren, Kaiqi Huang, and Tieniu Tan. Weakly Supervised Object Localization with Latent Category Learning. In *European Conference on Computer Vision (ECCV)*, 2014. 3

[30] Hakan Bilen, Marco Pedersoli, and Tinne Tuytelaars. Weakly supervised object detection with convex clustering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 3

[31] Xinggang Wang, Zhuotun Zhu, Cong Yao, and Xiang Bai. Relaxed Multiple-Instance SVM with Application to Object Discovery. In *IEEE International Conference on Computer Vision (ICCV)*, 2015. 3

[32] Ce Ge and Jingyu Wang. Fewer is More : Image Segmentation Based Weakly Supervised Object Detection with Partial Aggregation. In *The British Machine Vision Conference (BMVC)*, 2018. 3

[33] Dong Li, Jia-Bin Huang, Yali Li, Shengjin Wang, and Ming-Hsuan Yang. Weakly Supervised Object Localization with Progressive Domain Adaptation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3

[34] Zequn Jie, Yunchao Wei, Xiaojie Jin, Jiashi Feng, and Wei Liu. Deep Self-Taught Learning for Weakly Supervised Object Localization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 3

[35] Peng Tang, Xinggang Wang, Xiang Bai, and Wenyu Liu. Multiple Instance Detection Network with Online Instance Classifier Refinement. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 3, 4

[36] Ke Yang, Dongsheng Li, and Yong Dou. Towards Precise End-to-end Weakly Supervised Object Detection Network. In *IEEE International Conference on Computer Vision (ICCV)*, 2019. 3

[37] Yunhang Shen, Rongrong Ji, Zhiwei Chen, Yongjian Wu, and Feiyue Huang. UWSOD: Toward Fully-Supervised-Level Capacity Weakly Supervised Object Detection. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020. 3

[38] Peng Tang, Xinggang Wang, Song Bai, Wei Shen, Xiang Bai, Wenyu Liu, and Alan Yuille. PCL: Proposal Cluster Learning for Weakly Supervised Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2018. 3

[39] Satoshi Kosugi, Toshihiko Yamasaki, and Kiyoharu Aizawa. Object-Aware Instance Labeling for Weakly Supervised Object Detection. In *IEEE International Conference on Computer Vision (ICCV)*, 2019. 3

[40] Xiaopeng Zhang, Jiashi Feng, Hongkai Xiong, and Qi Tian. Zigzag Learning for Weakly Supervised Object Detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[41] Chenhao Lin, Siwen Wang, Dongqi Xu, Yu Lu, and Wayne Zhang. Object Instance Mining for Weakly Supervised Object Detection. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2020. 3

[42] Zhongzheng Ren, Zhiding Yu, Xiaodong Yang, Ming-Yu Liu, Yong Jae Lee, Alexander G. Schwing, and Jan Kautz. Instance-aware, Context-focused, and Memory-efficient Weakly Supervised Object Detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[43] Gong Cheng, Junyu Yang, Decheng Gao, Lei Guo, and Junwei Han. High-Quality Proposals for Weakly Supervised Object Detection. *IEEE Transactions on Image Processing (TIP)*, 2020. 3

[44] Yao Li, Linqiao Liu, Chunhua Shen, and Anton van den Hengel. Image Co-localization by Mimicking a Good Detector's Confidence Score Distribution. In *European Conference on Computer Vision (ECCV)*, 2016. 3

[45] Fang Wan, Pengxu Wei, Jianbin Jiao, Zhenjun Han, and Qixiang Ye. Min-Entropy Latent Model for Weakly Supervised Object Detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3, 7

[46] Yunhang Shen, Rongrong Ji, Changhu Wang, Xi Li, and Xuelong Li. Weakly Supervised Object Detection via Object-Specific Pixel Gradient. *IEEE Transactions on Neural Networks and Learning Systems (TNNLS)*, 2018. 3

[47] Yunhang Shen, Rongrong Ji, Kuiyuan Yang, Cheng Deng, and Changhu Wang. Category-Aware Spatial Constraint for Weakly Supervised Detection. *IEEE Transactions on Image Processing (TIP)*, 2019. 3

[48] Fang Wan, Chang Liu, Wei Ke, Xiangyang Ji, Jianbin Jiao, and Qixiang Ye. C-MIL: Continuation Multiple Instance Learning for Weakly Supervised Object Detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[49] Aditya Arun, C. V. Jawahar, and M. Pawan Kumar. Dissimilarity Coefficient based Weakly Supervised Object Detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[50] Xiaopeng Zhang, Yang Yang, and Jiashi Feng. Learning to Localize Objects with Noisy Labeled Instances. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2019. 3

[51] Boxiao Liu, Yan Gao, Nan Guo, Xiaochun Ye, Haihang You, and Dongrui Fan. Utilizing the Instability in Weakly Supervised Object Detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2019. 3

[52] Yunhan Shen, Rongrong Ji, Shengchuan Zhang, Wangmeng Zuo, and Yan Wang. Generative Adversarial Learning Towards Fast Weakly Supervised Detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[53] Ze Chen, Zhihang Fu, Rongxin Jiang, Yaowu Chen, and Xian-sheng Hua. SLV : Spatial Likelihood Voting for Weakly Supervised Object Detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[54] Ke Yang and Zhiyuan Wang. Objectness Consistent Representation for Weakly Supervised Object Detection. In *ACMMM*, 2020. 3

[55] Ke Yang, Peng Qiao, Zhiyuan Wang, Tianlong Shen, and Dongsheng Li. Rethinking Segmentation Guidance for Weakly Supervised Object Detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2020. 3

[56] Yunhang Shen, Rongrong Ji, Yan Wang, Zhiwei Chen, Feng Zheng, Feiyue Huang, and Yunsheng Wu. Enabling Deep Residual Networks for Weakly Supervised Object Detection. In *European Conference on Computer Vision (ECCV)*, 2020. 3, 6, 8

[57] Linpu Fang, Hang Xu, Zhili Liu, Sarah Parisot, and Zhenguo Li. EHSOD: CAM-Guided End-to-end Hybrid-Supervised Object Detection with Cascade Refinement. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2020. 3

[58] Mengmeng Xu, Yancheng Bai, and Bernard Ghanem. Missing Labels in Object Detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2019. 3

[59] Yunhang Shen, Rongrong Ji, Zhiwei Chen, Xiaopeng Hong, Feng Zheng, Jianzhuang Liu, Mingliang Xu, and Qi Tian. Noise-Aware Fully Webly Supervised Object Detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[60] Yunchao Wei, Zhiqiang Shen, Bowen Cheng, Honghui Shi, Jinjun Xiong, Jiashi Feng, and Thomas Huang. TS2C: Tight Box Mining with Surrounding Segmentation Context for Weakly Supervised Object Detection. In *European Conference on Computer Vision (ECCV)*, 2018. 3

[61] Xiaoyan Li, Meina Kan, Shiguang Shan, and Xilin Chen. Weakly Supervised Object Detection with Segmentation Collaboration. In *IEEE International Conference on Computer Vision (ICCV)*, 2019. 3

[62] Yan Gao, Liu Boxiao, Guo Nan, Ye Xiaochun, Wan Fang, You Haihang, and Fan Dongrui. C-MIDN : Coupled Multiple Instance Detection Network with Segmentation Guidance for Weakly Supervised Object Detection. In *IEEE International Conference on Computer Vision (ICCV)*, 2019. 3

[63] Jiwoon Ahn and Suha Kwak. Learning Pixel-level Semantic Affinity with Image-level Supervision for Weakly Supervised Semantic Segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[64] Ross Girshick. Fast R-CNN. In *IEEE International Conference on Computer Vision (ICCV)*, 2015. 3, 4

[65] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *IEEE International Conference on Computer Vision (ICCV)*, 2017. 3, 4, 7

[66] J. R. R Uijlings, K. E. A van de Sande, T Gevers, and A. W. M. Smeulders. Selective Search for Object Recognition. *International Journal of Computer Vision (IJCV)*, 2013. 4

[67] P Arbeláez, J Pont-Tuset, J Barron, F Marques, and J Malik. Multiscale Combinatorial Grouping. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 4, 6

[68] Jaume Amores. Multiple instance classification: Review, taxonomy and comparative study. *Artificial Intelligence (AI)*, 2013. 4

[69] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. "GrabCut": interactive foreground extraction using iterated graph cuts. In *ACM SIGGRAPH*, 2004. 5

[70] Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. 2004. 5

[71] J Deng, W Dong, R Socher, L.-J. Li, K Li, and L Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 6

[72] Yi Zhu, Yanzhao Zhou, Qixiang Ye, Qiang Qiu, and Jianbin Jiao. Soft Proposal Networks for Weakly Supervised Object Localization. In *IEEE International Conference on Computer Vision (ICCV)*, 2017. 7

[73] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 8