Toward Joint Thing-and-Stuff Mining for Weakly Supervised Panoptic Segmentation

Yunhang Shen[1], Liujuan Cao[1]*, Zhiwei Chen[1], Feihong Lian[1], Baochang Zhang[2]
Chi Su[3], Yongjian Wu[4], Feiyue Huang[4], Rongrong Ji[1,5,6]
[1]Media Analytics and Computing Lab, Department of Artificial Intelligence,
School of Informatics, Xiamen University, 361005, China, [2]Institute of Artificial Intelligence,
Beihang University, Beijing, China, [3]KingSoft Cloud Co. Ltd., Beijing, China
[4]Tencent Youtu Lab, Shanghai, China, [5]Institute of Artificial Intelligence,
Xiamen University, Xiamen, China, [6]Peng Cheng Laborotory, Shenzhen, China
shenyunhang01@gmail.com, caoliujuan@xmu.edu.cn, zhiweichen.xmu@gmail.com
lianfh@stu.xmu.edu.cn, bczhang@buaa.edu.cn, suchi@kingsoft.com
littlekenwu@tencent.com, garyhuang@tencent.com, rrji@xmu.edu.cn

Abstract

Panoptic segmentation aims to partition an image to object instances and semantic content for thing and stuff categories, respectively. To date, learning weakly supervised panoptic segmentation (WSPS) with only image-level labels remains unexplored. In this paper, we propose an efficient jointly thing-and-stuff mining (JTSM) framework for WSPS. To this end, we design a novel mask of interest pooling (MoIPool) to extract fixed-size pixel-accurate feature maps of arbitrary-shape segmentations. MoIPool enables a panoptic mining branch to leverage multiple instance learning (MIL) to recognize things and stuff segmentation in a unified manner. We further refine segmentation masks with parallel instance and semantic segmentation branches via self-training, which collaborates the mined masks from panoptic mining with bottom-up object evidence as pseudo-ground-truth labels to improve spatial coherence and contour localization. Experimental results demonstrate the effectiveness of JTSM on PASCAL VOC and MS COCO. As a by-product, we achieve competitive results for weakly supervised object detection and instance segmentation. This work is a first step towards tackling challenge panoptic segmentation task with only image-level labels.

1. Introduction

Panoptic segmentation focuses on simultaneously segmenting all object instances and semantic content in an image. It is one of the most important tasks in computer vision due to its great academic values and industrial applications. Recent rapid progress on panoptic segmentation has been driven by combining the strength of instance segmentation and semantic segmentation tasks via a multi-branch scheme. However, these deep models heavily rely on a large amount of training data with expensive instance-level and pixel-wise annotations. Collecting such training data has been a particular bottleneck on the way of applying panoptic segmentation to real-world applications, *e.g.*, autonomous driving, robotics, and image editing, where labelling each pixel for numerous images is particularly time-consuming. For example, fully annotating a single image in Cityscapes [1] required more than 1.5 hours on average.

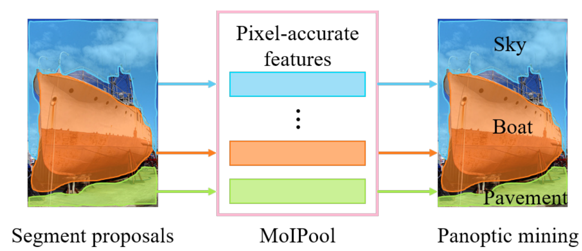

Figure 1: The overall flowchart of our JTSM framework.

One way to reduce the requirement of strong supervision is the weakly supervised panoptic segmentation (WSPS), which seeks to use weak annotations for model training. To our best knowledge, the only previous work that attempted to address WSPS problem is that of [2], which requires bounding boxes for *thing* categories and image-level tags for *stuff* during training. However, for applications needing very large-scale image sets and categories, bounding-box-level annotations still require enormous human effort. It is

*Corresponding author.

thus desirable to learn panoptic segmentation from large-scale datasets with weaker supervision.

We focus on the most extreme case of WSPS where only image-level labels are available, and no instance-level annotations are involved during training. To date, none of the existing work further investigates the problem of learning panoptic segmentation with only image-level labels. An intuitive and strong baseline method is to perform weakly supervised instance segmentation (WSIS) and weakly supervised semantic segmentation (WSSS) independently, and use heuristic post-processing method [3] to merge their results. However, the straightforward combination of such two techniques disregards the underlying relationship and fails to borrow rich contextual cues between *things* and *stuff*. As context information is critical to recognize and localize the objects, and foreground objects provide complementary cues to assist background understanding [4, 5].

In this paper, we propose a Joint *Thing*-and-*Stuff* Mining (JTSM) framework to learn panoptic segmentation with only image-level labels, as illustrated in Fig. 1. Our motivation is to consider foreground *things* and background *stuff* as uniform object instances in form of segmentation masks. Particularly, each connected component of *stuff* content is viewed as an individual instance, which shares the same spirit as *thing* objects. Different to the baseline that frames the two related tasks at architectural level via a multi-branch scheme, the main advantage of JMTS is to model the correlations between objects and background at instance level.

To this end, we design a novel mask of interest pooling (MoIPool) to extract fixed-size pixel-accurate feature maps for arbitrary-shape segmentations, which provides a uniform representational power for *things* and *stuff*. Thus, given a set of segment proposals, a panoptic mining branch leverages multiple instance learning (MIL) to mine all target categories in a unified manner. We further introduce two schemes to refine segmentation masks. First, we collaborate the mined results from panoptic mining with bottom-up object evidence to improve spatial coherence and contour localization. Second, we introduce self-training to refine segmentation masks with parallel instance and semantic segmentation branches. With pseudo-ground-truth masks from its preceding branch, the discriminatory power of the image segmentation can be enhanced. Experimental results demonstrate the effectiveness of our proposed JTSM compared to strong baselines on PASCAL VOC [6] and MS COCO [7]. As a by-product, we also achieve competitive results for both weakly supervised object detection and instance segmentation tasks.

The contributions of this work are three folds:

- We propose JTSM to jointly segment *things* and *stuff* for weakly supervised panoptic segmentation in a unified framework. To our best knowledge, this work makes the first attempt to tackle challenge panoptic segmentation task with only image-level labels.
- We design a novel mask of interest pooling (MoIPool) to compute fixed-size pixel-accurate feature maps of arbitrary-shape segmentations, which enables JTSM to leverage multiple instance learning (MIL) to mine *thing* and *stuff* with a uniform representational power.
- Self-training is further introduced to refine the image segmentation with two parallel instance and semantic segmentation branches, which are supervised by the mined results and bottom-up object evidence to improve spatial coherence and contour localization.

2. Related Work

Weakly Supervised Panoptic Segmentation (WSPS). Although learning panoptic segmentation with only image-level labels is challenging without any existing work in previous literature, one attempt with bounding-box-level annotations has been made [2]. Our work has significant differences to [2]. First, method in [2] required bounding-box supervision and fully-labelled examples of some categories. We use only image-level labels to learn panoptic segmentation for the first time. Second, they [2] heavily relied on external models to pre-compute pseudo-ground-truth masks for *thing* and *stuff* categories independently, which failed to model the intrinsic interaction between semantic segmentation and instance segmentation. However, our method simultaneously segments all target categories in a unified manner to achieve a holistic understanding of an image.

Weakly Supervised Object Detection (WSOD). WSOD aims to predict object instance in the form of bounding boxes with weak supervision. Recent widely-used WSOD alternates between localizing object instances and training appearance representation via multiple instance learning (MIL). For example, WSDDN [8] selected box proposals by parallel detection and classification branches in deep convolutional neural networks (CNNs). This method is extended by leveraging contextual information [9], gradient map [10, 11], attention mechanism [12], semantic segmentation [13, 14, 15] to suppress low-quality box proposals. Some work in [16, 17, 18, 19] treated the top-scoring proposals as supervision to train multiple instance refinement classifiers. Other different strategies [20, 21, 22, 23, 24, 25, 26] were also proposed to generate pseudo-ground-truth boxes and assign labels to box proposals. And the above framework was further improved by min-entropy prior [27, 28], continuation MIL [29], utilizing uncertainty [30, 31, 32], knowledge distillation [33], spatial likelihood voting [34], objectness consistent [35, 36] and generative adversarial learning [37]. Methods in [38, 39, 40] trained object detection systems from different supervisions.

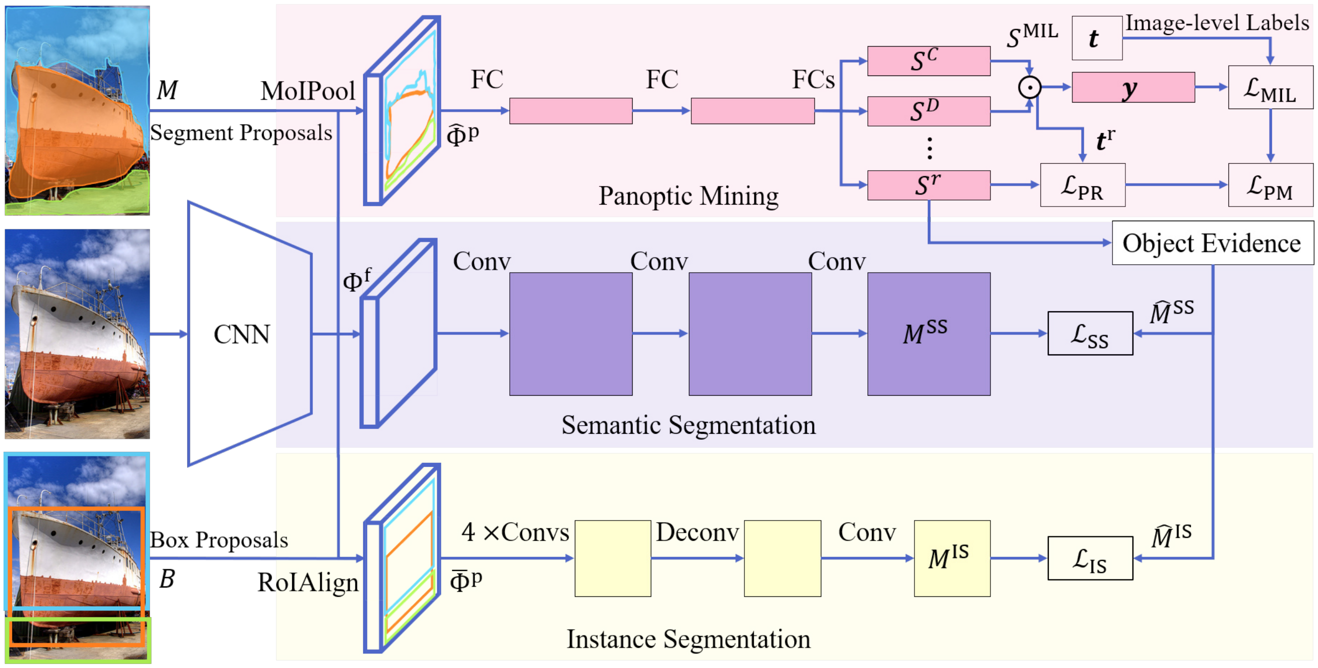

Figure 2: The figure illustrates the overall architecture of JTSM. Given an image, panoptic mining branch jointly segments *things* and *stuff* in a unified manner, which are refined by instance segmentation and semantic segmentation, respectively.

Weakly Supervised Instance Segmentation (WSIS). WSIS methods can be categorized into two groups. The first group utilizes bounding-box annotations as weak supervision to training WSIS models. Most methods in this group used box-driven segmentation [41, 42] or multiple instance learning [43] to generate instance-level pseudo-ground-truth labels, which are then refined by recursive training [41, 2]. The second group further challenges WSIS problem with only image-level supervision. The early work [44, 45] utilized class response maps to capture visual cues via back-propagation, which are used to generate instance masks from object segment proposals. WISE [46] and IRNet [47] generated coarse masks from class activation maps [48], which is regarded as pseudo-ground-truth labels to train fully supervised models. S^{4}Net [49] and LIID [50] further leveraged graph partitioning algorithms to learn pseudo-ground-truth labels. Label-PEnet [51] transform image-level labels to pixel-wise predictions with multiple cascaded modules and curriculum learning strategy. Kim *et al*. [52] proposed multi-task community learning to construct a positive feedback loop and generates pseudo-ground-truths masks using class activation maps [48].

Weakly Supervised Semantic Segmentation (WSSS). Recently, lots of WSSS methods have been proposed to alleviate labelling cost. Many early work [53, 54, 55, 44] leveraged CNN built-in pixel-level cues and constraint priors to learn segmentation masks. Pathak *et al.* [53] 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. Some works derive category-wise saliency maps from intermediate feature maps of CNNs to estimate the segmentation masks [54, 44]. Recently, WSSS methods [56, 57, 58, 59] often treat initial object localization cues as pseudo supervision and train fully supervised segmentation models. Popular methods [60, 61, 49, 62, 63] leveraged object saliency maps and feature activation maps to provide complimentary information. Many regularizations [56, 64, 65, 66, 63] were proposed to improve the segmentation results. There are also works [67, 68, 69, 70, 71] that focused on improving feature learning in iterative frameworks. Various approaches based on iteratively mining common feature [72, 73], region refinement [59, 74], random-walk label propagation [75], dilated convolution [57] and pixel-level semantic affinity [58] were proposed. Work in [73, 76] also explored object boundaries to refine localization maps.

3. The Proposed Method

3.1. Overall Framework

The overview of our proposed Joint *Thing*-and-*Stuff* Mining (JTSM) framework is illustrated in Fig. 2 We construct a parallel multi-branch architecture for panoptic mining, instance segmentation and semantic segmentation, respectively. Each branch takes full-image feature maps from the backbone network as input. First, panoptic mining branch leverages multiple-instance learning (MIL) [77] to jointly segment *thing* and *stuff* object with multiple panoptic refinement heads. Particularly, we design a novel MoIPool to produce fixed-size pixel-accurate convolutional feature maps for segment proposals, which are generated by unsupervised proposal generation methods [78, 79], Second, mined masks from panoptic mining are integrated with bottom-up object evidence to improve spatial coherence and contour localization. Third, parallel instance and semantic segmentation branches further refine *thing* and *stuff* masks by taking the predictions as supervision. During training, we have the following objective function

$$\mathcal{L} = \mathcal{L}_{\mathrm{PM}} + \mathcal{L}_{\mathrm{IS}} + \mathcal{L}_{\mathrm{SS}}, \tag{1}$$

where $\mathcal{L}_{\mathrm{PM}}$ is the loss functions of panoptic mining branch, $\mathcal{L}_{\mathrm{IS}}$ and $\mathcal{L}_{\mathrm{SS}}$ are the loss functions for instance and semantic segmentation branches, respectively.

3.2. Joint Thing-and-Stuff Mining

The panoptic mining branch aims to jointly segment countable *thing* instances and uncountable *stuff* content in a unified manner. Recall that background *stuff* can be partitioned into a set of connected components. Thus, we consider each connected component of background as an individual instance, which shares the same spirit as countable objects. Although distinguishing disconnected components for background is unnecessary, all *things* and *stuff* are viewed as uniform object instances. To this end, we follow the MIL pipeline in deep convolutional networks and convert two-stream WSDDN [8] and OICR [18] algorithms to recognize instances for all categories in a unified manner.

Formally, given an image I and corresponding image-level labels $\mathbf{t} = [\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_{n^c}]$ during training, JTSM aims to estimate segmentation mask for each object instance in this image. Let $\mathbf{t}$ be a fixed-length 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 total number of *thing* and *stuff* categories. The backbone network first outputs full-image feature maps Φ^{f} of input I. Then we use MoIPool layer (discussed later) to compute fixed-size pooled feature maps Φ^{p} for segment proposals, which are followed by two fully-connected layers with ReLu activation and dropout layer to extract final proposal features. After that, a MIL head forks the proposal features into two streams to produce two score matrices $S^{\mathrm{C}}, S^{\mathrm{D}} \in \mathbb{R}^{n^{\mathrm{p}} \times n^{\mathrm{c}}}$ by another two fully-connected layers, respectively, where n^{p} is the number of proposals. Finally, we use the element-wise product to compute the final proposal score matrix as $S^{\mathrm{MIL}} = \sigma(S^{\mathrm{C}}) \odot \sigma((S^{\mathrm{D}})^T)^T$, where $\sigma(\cdot)$ is the softmax function. To train the MIL head with only image-level supervision, a sum pooling is applied to acquire image-level multi-label classification scores as $\mathbf{y}_c = \sum_{p=1}^{n^{\mathrm{p}}} S^{\mathrm{MIL}}_{pc}$. Then we obtain a multi-label cross-entropy objective function

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

To further reduce mis-recognize, we refine MIL scores via multiple panoptic refinement heads, each of which contains a single fully-connected layer. For the r^{th} refinement head, it reuses proposal features as input and produces new classification scores $S^{\mathrm{r}} \in \Re^{n^{\mathrm{p}} \times (n^{\mathrm{c}}+1)}$, where $n^{\mathrm{c}} + 1$ indicates the 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 $S^{\mathrm{r}-1}$ is selected as pseudo-ground-truth labels and assigns positive/negative labels for the rest segment proposals. We also set $S^0 = S^{\mathrm{MIL}}$. Thus, the corresponding panoptic refinement loss is

$$\mathcal{L}^{\mathrm{r}}_{\mathrm{PR}} = -\sum_{p=1}^{n^{\mathrm{p}}} \mathbf{y}_{\mathbf{t}^{\mathrm{r}}_p} \log \Big(\frac{\exp(S^{\mathrm{r}}_{p\mathbf{t}^{\mathrm{r}}_p})}{\sum_j \log(S^{\mathrm{r}}_{pj})} \Big), \tag{3}$$

where $\mathbf{t}^{\mathrm{r}}_{\mathrm{p}}$ denotes the classification targets for the p^{th} segment proposal in the r^{th} head, and $S^{\mathrm{r}}_{p\mathbf{t}^{\mathrm{r}}_p}$ is the corresponding prediction score. Thus, L_{PR} is the softmax cross-entropy loss weighted by image-level classification scores $\mathbf{y}_{\mathbf{t}^{\mathrm{r}}_p}$.

With above definitions, the overall objective function for panoptic mining branch is defined as

$$\mathcal{L}_{\mathrm{PM}} = \mathcal{L}_{\mathrm{MIL}} + \sum_{r=1}^{n^{\mathrm{r}}} \mathcal{L}^{\mathrm{r}}_{\mathrm{PR}}, \tag{4}$$

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

3.3. Mask of Interest Pooling (MoIPool)

We design a novel Mask of Interest Pooling to compute fixed-size feature maps for segment proposals. Different to RoIPool [80] and RoIAlign [81] that require rectangle proposals, *i.e.*, bounding boxes, MoIPool enables to extract pixel-accurate feature maps of arbitrary-shape segmentations. To this end, we introduce two efficient variants: shape-interpolation and shape-invariant MoIPool.

The first variant is shape-interpolation MoIPool, which only applies the pooling operation inside the segment proposals. Our intuition is that if non-rigid segmentations are transformed into rigid regions, we can reuse traditional methods, *e.g.*, RoIPool and RoIAlign. Therefore, we first interpolate segmentations to rectangle regions by thin plate splines (TPS) algorithm, which has been widely used as the non-rigid transformation model in image alignment and shape matching. TPS produces smooth surfaces, which are infinitely differentiable. Despite its simplicity, experiment results show that shape-interpolation MoIPool achieves competitive performance compared to shape-invariant one.

We further design a shape-invariant MoIPool to maintain accurate contour information of segment proposals. Suppose that the backbone network extracts a full-image feature map $\Phi^f \in \Re^{h^\mathrm{f} \times w^\mathrm{f}}$, which has a total stride size s. We omit the channels of feature maps for simplification. Each segment proposal is defined by a binary mask M, which has the same spatial size as the input image I. We can easily obtain the corresponding bounding boxes B of segment proposal, which is defined by a four-tuple $(x^\mathrm{b}, y^\mathrm{b}, w^\mathrm{b}, h^\mathrm{b})$ that specifies its top-left corner $(y^\mathrm{b}, x^\mathrm{b})$ and its width and height $(w^\mathrm{b}, h^\mathrm{b})$. We also denote the pooled proposal feature map as $\Phi^p \in \Re^{h^\mathrm{p} \times w^\mathrm{p}}$, where $h^\mathrm{p} \times w^\mathrm{p}$ is the pre-difined spatial size of pooled features. The proposed MoIPool works by dividing the $h^\mathrm{b}/s \times w^\mathrm{b}/s$ cropped proposal feature map into a $h^\mathrm{p} \times w^\mathrm{p}$ grid of sub-windows of approximate size $h^\mathrm{b}/s/h^\mathrm{p} \times w^\mathrm{b}/s/w^\mathrm{p}$. Maximum value in each sub-window is assigned into the corresponding output grid cell as

$$\begin{gathered}\Phi^p_{uv} = \max(\Omega_{ij}\Phi^f_{ij}),\\ i \in \left[y^\mathrm{b}/s + \lfloor h^\mathrm{b}/s \cdot u/h^\mathrm{p}\rfloor, y^\mathrm{b}/s + \lceil h^\mathrm{b}/s \cdot (u+1)/h^\mathrm{p}\rceil\right],\\ j \in \left[x^\mathrm{b}/s + \lfloor w^\mathrm{b}/s \cdot v/w^\mathrm{p}\rfloor, x^\mathrm{b}/s + \lceil w^\mathrm{b}/s \cdot (v+1)/w^\mathrm{p}\rceil\right],\end{gathered} \tag{5}$$

where $\Omega \in \Re^{h^\mathrm{f} \times w^\mathrm{f}}$ is an indicator matrix, which equals 1 if the corresponding elements in proposal feature maps are available for max-pooling. Given the i^th row and the j^th column element in proposal feature maps, we crop the corresponding sub-window in binary mask $M \in \Re^{h^\mathrm{m} \times w^\mathrm{m}}$ of segment proposals, and acquire maximum values as

$$\begin{gathered}\Omega_{uv} = \max M_{ij}\\ i \in [\lfloor u \cdot s\rfloor, \lceil (u+1)\cdot s\rceil], j \in [\lfloor v \cdot s\rfloor, \lceil (v+1)\cdot s\rceil].\end{gathered} \tag{6}$$

However, the above definition fills activations to zeros if corresponding sub-windows do not belong to segment proposals. To align the feature activations among different proposals, we further introduce a compensation term

$$\psi = \frac{h^\mathrm{p} w^\mathrm{p}}{\sum_{ij} \Phi^\mathrm{p}}. \tag{7}$$

Thus, the final pooled proposal features are scaled up as

$$\hat{\Phi}^p = \psi \Phi^\mathrm{p}. \tag{8}$$

In fact, the proposed MoIPool can be regarded as a generalization of RoIPool. As MoIPool degenerates to RoIPool when the segment proposal is a rectangle window.

3.4. Segmentation Refinement

As the quality of segmentation results from panoptic mining branch heavily rely on segment proposals, we further take the advantage of self-train to refine masks. To do this we introduce two parallel instance and semantic segmentation branches, which are supervised by pseudo-ground-truth masks generated from panoptic mining.

In details, instance segmentation branch consists of 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 a final prediction layer with 1×1 kernels. Instance segmentation takes proposal features $\bar{\Phi}^\mathrm{p}$ from RoIAlign [81] as input and produces refined masks M^IS for all n^t *thing* categories. Thus, given a set of box proposals, instance segmentation objective function is

$$\mathcal{L}_\mathrm{IS} = \sum_{p=1}^{n^\mathrm{p}} \sum_{c=1}^{n^\mathrm{t}} [\mathbf{t}^\mathrm{IS}_p = c] \mathbf{y}_{\mathbf{t}^\mathrm{IS}_p} \mathcal{L}_\mathrm{BCE}(M^\mathrm{IS}_{pc}, \hat{M}^\mathrm{IS}_{pc}), \tag{9}$$

where M^IS_{pc} and $\hat{M}^\mathrm{IS}_p$ are the predicted and target masks for the p^th proposals and the c^th category, and $\mathbf{t}^\mathrm{IS}_p$ is the pseudo category labels for the p^th proposals. And $\mathcal{L}_\mathrm{BCE}$ is the binary cross-entropy loss. As mask head is class-specific, we only compute losses for categories existed in images, which are then weighted by the image-level prediction scores $\mathbf{y}_{\mathbf{t}^\mathrm{IS}_p}$.

Semantic segmentation branch consists of two convolutional layers with 3×3 kernels and 256 channels to extract feature maps and a final prediction layer with 1×1 kernels. Different to instance segmentation, semantic segmentation branch takes full-image feature maps Φ^f as input and outputs refined masks M^SS for each *stuff* category. Thus, the semantic segmentation objective function is defined as

$$\mathcal{L}_\mathrm{SS} = \mathcal{L}_\mathrm{CE}(M^\mathrm{SS}, \hat{M}^\mathrm{SS}), \tag{10}$$

where $\hat{M}^\mathrm{SS}$ denotes the target segmentation masks, and $\mathcal{L}_\mathrm{BCE}$ is the binary cross entropy loss function.

To generate pixel-wise supervision $\hat{M}^\mathrm{IS}$ and $\hat{M}^\mathrm{SS}$ for segmentation refinement, we integrate the mined masks of panoptic mining branch with bottom-up object evidence to improve spatial coherence and contour localization. We employ unsupervised grouping-based segmentation GrabCut [86] algorithm to re-estimate object masks within the corresponding bounding boxes. For each *thing* and *stuff* category existed in images, we constrain the area within the highest-score segment proposal as firm positive pixels and the areas outside its bounding boxes as firm negative pixels during re-estimation. Such post-process of mined masks helps to reduce ambiguous outline using low-level features such as the pixel colours. We are not restricted with the algorithms that generate object evidence from input images. The re-estimated masks are treated as pseudo-ground-truth masks to learn above segmentation refinement.

Table 1: Ablation study of different proposal-pooling methods on PASCAL VOC 2012 panoptic segmentation.

Methods	PQ	SQ	RQ	PQ^{Th}	SQ^{Th}	RQ^{Th}	PQ^{St}	SQ^{St}	RQ^{St}
RoIPool	36.5	74.4	48.0	34.6	73.3	46.0	75.6	84.6	90.3
RoIAlgin	36.2	74.2	47.7	34.4	73.2	43.8	74.3	84.3	89.1
MoIPool									
Shape-interpolation	37.2	74.2	48.8	35.3	73.2	46.9	76.0	84.4	90.9
Shape-invariant	39.0	74.4	51.5	37.1	73.9	49.5	77.7	85.1	91.2

Table 2: Ablation study of segmentation refinement on PASCAL VOC 2012 panoptic segmentation.

$\mathcal{L}_{\text{IS}}$	$\mathcal{L}_{\text{SS}}$	Re-estimate	PQ	SQ	RQ	PQ^{Th}	SQ^{Th}	RQ^{Th}	PQ^{St}	SQ^{St}	RQ^{St}
			30.7	73.4	41.4	29.0	72.9	39.5	66.9	81.8	81.8
✓			34.9	73.6	46.2	33.9	73.1	44.9	69.4	81.7	85.1
	✓		33.9	73.2	45.5	28.8	72.7	39.2	76.6	83.5	91.6
✓	✓		36.5	73.9	48.6	34.6	73.4	46.5	76.7	84.0	91.1
✓	✓	✓	39.0	74.4	51.5	37.1	73.9	49.5	77.7	85.1	91.2

Table 3: Ablation study of different n^{r} values in Equ. 4 on PASCAL VOC 2012 panoptic segmentation.

n^{r}	PQ	SQ	RQ	PQ^{Th}	SQ^{Th}	RQ^{Th}	PQ^{St}	SQ^{St}	RQ^{St}
0	30.3	71.9	41.5	28.3	71.3	39.3	71.4	82.0	87.1
1	35.0	72.2	47.8	33.6	69.8	44.9	73.2	82.7	89.2
2	36.5	72.8	50.3	36.1	72.2	47.5	75.3	83.8	89.4
3	38.1	73.9	50.8	36.5	73.0	48.8	76.1	84.6	90.2
4	39.0	74.4	51.5	37.1	73.9	49.5	77.7	85.1	91.2
5	38.6	74.3	51.5	36.7	73.6	49.8	77.8	85.6	91.0

4. Quantitative Evaluations

4.1. Experimental Setup

Datasets We evaluate our method on two popular benchmarks, *i.e*., PASCAL VOC 2012 [6] and MS COCO [7]. PASCAL VOC 2012 consists of 20 target categories as well as one background category. As in full supervised panoptic segmentation [87, 82], we generate a *training* set by merging the Pascal VOC 2012 *training* set and the additional annotations from the SBD dataset [88]. This results in $10,582$ training images. For validation, we evaluate on the Pascal VOC 2012 *validation* set, as the evaluation server is not available for panoptic segmentation. The MS COCO panoptic segmentation has a greater number of images and categories. It features 118k training images, 5k validation images. There are 133 semantic classes, including 53 *stuff* and 80 *thing* categories. We also evaluate the performance of object detection on PASCAL VOC 2007 [6], which are widely-used benchmark dataset for WSOD. PASCAL VOC 2007 consists of $5,011$ *trainval* images, and $4,092$ *test* images over 20 categories. Note that only image-level labels are used for model training in all our results.

Evaluation Protocol. Our main evaluation metric is the panoptic quality (PQ), which is the product of segmentation quality (SQ) and recognition quality (RQ) [3]. SQ captures the average segmentation quality of matched segments, whereas RQ measures the ability of an algorithm to detect objects correctly. For the evaluation metrics of instance segmentation, we also report the standard MS COCO metrics [7], which is mean average precision (AP) over IoU thresholds. For object detection on Pascal VOC, we follow 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. For object detection on MS COCO, we report standard COCO metrics, including AP at different IoU thresholds.

Implementation Details We implement our method using PyTorch framework. All backbones are initialized with the weights pre-trained on ImageNet ILSVRC [89]. 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 $4\times$ training schedules for MS COCO. In the multi-scale setting, we use scales range from 480 to $1,216$ 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. We use MCG [79] to generate segment proposals for all experiments. We set the maximum number of proposals in an image to be $4,000$. The test scores are the average of scales of $\{480, 576, 688, 864, 1200\}$ and flips. Detection results are post-processed by NMS with a threshold of 0.5.

We use the following parameter settings in all the experiments unless specified otherwise. We set the number n^{r} of object refinement branches to 4. For the proposed MoIPool, we use the shape-invariant version for default.

4.2. Weakly Supervised Panoptic Segmentation

We first 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 panoptic segmentation as described above. Here, we use ResNet18-WS [90] as the backbone to save

Table 4: Comparison with the state-of-the-art methods on PASCAL VOC 2012 panoptic segmentation. The terms $\mathcal{M}$, $\mathcal{B}$ and $\mathcal{I}$ denote pixel-level, bounding-box-level and image-level labels, respectively

Method	Supervision	Backbone	PQ	SQ	RQ	PQ^{Th}	SQ^{Th}	RQ^{Th}	PQ^{St}	SQ^{St}	RQ^{St}
DeeperLab [82]	$\mathcal{M}$	Xception-71	67.4	-	-	-	-	-	-	-	-
Panoptic FPN [83]	$\mathcal{M}$	ResNet50	65.7	84.3	77.6	64.5	83.9	76.5	90.8	92.5	98.1
Li *et al.* [2]	$\mathcal{B}+\mathcal{I}$	ResNet101	59.0	-	-	-	-	-	-	-	-
Combination [47, 58]	$\mathcal{I}$	ResNet50	37.1	69.8	49.5	35.5	70.5	47.2	74.2	82.6	86.3
JTSM	$\mathcal{I}$	ResNet18-WS	**39.0**	**74.4**	**51.5**	**37.1**	**73.9**	**49.5**	**77.7**	**85.1**	**91.2**

Table 5: Comparison with the state-of-the-art methods on MS COCO panoptic segmentation.

Method	Supervision	Backbone	PQ	SQ	RQ	PQ^{Th}	SQ^{Th}	RQ^{Th}	PQ^{St}	SQ^{St}	RQ^{St}
Panoptic FPN [83]	$\mathcal{M}$	ResNet50	39.0	-	-	45.9	-	-	28.7	-	-
JTSM	$\mathcal{I}$	ResNet18-WS	5.3	30.8	7.8	8.4	46.6	11.4	0.7	6.4	0.5

Table 6: Comparison with the state-of-the-art methods on PASCAL VOC 2012 instance segmentation.

Method	Supervision	Backbone	$mAP_{0.50}$	$mAP_{0.75}$
Mask R-CNN [81]	$\mathcal{M}$	ResNet101	67.9	44.9
PRM [44]	$\mathcal{I}$	ResNet50	26.8	9.0
IAM [45]	$\mathcal{I}$	ResNet50	28.8	11.9
IRNet [47]	$\mathcal{I}$	ResNet50	46.7	-
Label-PEnet [51]	$\mathcal{I}$	VGG16	30.2	12.9
WISE [46]	$\mathcal{I}$	ResNet50	41.7	23.7
Kim *et al.* [52]	$\mathcal{I}$	ResNet50	35.7	5.8
Arun *et al.* [42]	$\mathcal{I}$	ResNet50	**50.9**	**28.5**
LIID [50]	$\mathcal{I}$	ResNet50	48.4	24.9
JTSM	$\mathcal{I}$	ResNet18-WS	44.2	12.0

time if not mentioned. When tuning each group of hyperparameters, other parameters are kept as default.

The number n^{r} of panoptic refinement heads. The panoptic refinement heads output final mining scores for segmentation during testing, which heavily influence the performance of instance and semantic segmentation. The hyperparameter n^{r} in Equ. 4 controls the number of panoptic refinement branches. Different settings and corresponding results of n^{r} are displayed in Tab. 3. When we have $n^{\mathrm{r}} = 0$, the second term of loss function $\mathcal{L}_{\mathrm{IR}}$ in Equ. 4 are omitted. We can see that the results of this setting are worse than using panoptic refinement branches, demonstrating that the panoptic refinement is very helpful for segmentation predictions. When $n^{\mathrm{r}} \geq 4$, the performance gains are margin. We use 4 as the default values for n^{r}.

The shape-invariant *vs*. shape-intepolation MoIPool. We first use the traditional RoIPool [80] and RoIAlgin [81] methods to analyze how performance varies with different proposal pooling methods. As shown in Tab. 1, traditional methods are unable to handle *stuff* well. As *stuff* content often has large outline and scale variance, which may also contain other *stuff* and *thing* object. Thus, it requires to pooling methods to compute pixel-accurate feature maps of arbitrary-shape regions. The proposed MoIPool achieves large performance gains compared to RoIPool and RoIAlgin, as MoIPool only utilizes the features within segment proposals. We also find that shape-invariant MoIPool has superior performance compared to shape-interpolation version. As shape-invariant MoIPool maintains accurate contour information of segment proposals.

The instance and semantic segmentation refinement. We continue by evaluating the effect of segmentation refinement. As shown in Tab. 2, segmentation refinement improve overall performance with large gains. As the quality of original mined masks heavily relies on segment proposals, while the segmentation refinement leverages self-training to improve predicted masks. We observe that the performance can increase significantly with the guidance of bottom-up evidence. It demonstrates that the bottom-up evidence is positively correlated to object segmentation.

With the above ablation study, we perform panoptic segmentation on the PASCAL VOC 2012 and MS COCO with various ResNet backbones. To the best of our knowledge, this is the first work reporting results for image-level supervised panoptic segmentation. Inspired by fully supervised panoptic segmentation, we construct a strong baseline for WSPS, which combines the output of independent WSIS and WSSS tasks via a series of post-processing steps [3] that merges their outputs. Specifically, we use the results from a combination of WSIS algorithm, IRNet [47], and WSSS algorithm, AffinityNet [58]. Note that both IRNet and AffinityNet are competitive approaches in their target tasks. For PASCAL VOC that has only one *stuff* category, we compute all *thing* segmentation and treat the rest regions as *stuff* segmentation. Tab. 4 and 5 show that JTSM significantly outperforms the strong baseline models that use the same setting, *i.e.*, using image-level labels only for model training. The performance improvement for *stuff*, *e.g.*, PQ^{St}, shows the validity of joint category mining, while the improvement for *thing*, *e.g.*, PQ^{Th}, indicates the effectiveness of MoIPool.

4.3. Weakly Supervised Instance Segmentation

We also report instance segmentation performance in terms of AP and compare with other WSIS methods on

Table 7: Comparison with the state-of-the-art methods on MS COCO instance segmentation.

Method	Supervision	Backbone	mAP	$m\text{AP}_{0.50}$	$m\text{AP}_{0.75}$	$m\text{AP}_S$	$m\text{AP}_M$	$m\text{AP}_L$
Mask R-CNN [81]	$\mathcal{M}$	ResNet101	35.7	58.0	37.8	15.5	38.1	52.4
WS-JDS [15]	$\mathcal{I}$	VGG16	**6.1**	11.7	**5.5**	**1.5**	**7.1**	12.2
JTSM	$\mathcal{I}$	ResNet18-WS	**6.1**	**12.1**	5.0	0.1	3.0	**12.6**

Table 8: Comparison with the state-of-the-art methods on PASCAL VOC 2007, 2012 and MS COCO object detection.

Method	Supervision	Bakcbone	PASCAL VOC 2007		PASCAL VOC 2012		MS COCO Avg. Precision, IoU:		
			mAP (%)	CorLoc (%)	mAP (%)	CorLoc (%)	0.5:0.95	0.5	0.75
Faster RCNN[84]	$\mathcal{B}$	VGG16	69.9	–	67.0	–	21.2	41.5	–
WSDDN [8]	$\mathcal{I}$	VGG16	34.8	53.5	–	–	9.5	19.2	8.2
OPG [10]	$\mathcal{I}$	VGG16	28.8	43.5	–	–	–	–	–
CSC C5 [11]	$\mathcal{I}$	VGG16	43.0	62.2	37.1	61.4	**12.9**	23.8	**13.2**
WS-JDS [15]	$\mathcal{I}$	VGG16	45.6	64.5	39.1	63.5	–	–	–
OICR [18]	$\mathcal{I}$	VGG16	41.2	60.6	37.9	62.1	–	–	–
MELM [28]	$\mathcal{I}$	VGG16	47.3	61.4	42.4	–	–	–	–
Kosugi *et al.* [21]	$\mathcal{I}$	VGG16	47.6	66.7	43.4	66.7	–	–	–
C-MIL [29]	$\mathcal{I}$	VGG16	50.5	65.0	46.7	67.4	–	–	–
Pred Net [30]	$\mathcal{I}$	VGG16	52.9	70.9	48.4	69.5	–	–	–
WSOD^2 [85]	$\mathcal{I}$	VGG16	53.6	69.5	47.2	71.9	10.8	22.7	–
Yanga *et al.* [19]	$\mathcal{I}$	VGG16	48.6	66.8	–	–	–	–	–
C-MIDN [14]	$\mathcal{I}$	VGG16	52.6	68.7	50.2	71.2	9.6	21.4	–
Ren *et al.* [24]	$\mathcal{I}$	VGG16	**54.9**	68.8	**52.1**	70.9	12.4	**25.8**	10.5
UWSOD [26]	$\mathcal{I}$	ResNet18-WS	45.0	63.8	46.2	65.7	3.1	10.1	1.4
JTSM	$\mathcal{I}$	ResNet18-WS	53.4	**71.4**	51.5	**72.5**	9.4	21.3	7.9

PASCAL VOC 2012. In details, JTSM only mines *thing* categories and ignores the segmantic branch. As shown in Tab. 6 and 7, our JTSM largely outperforms previous state-of-the-art that also uses image-level supervision. Some previous methods achieved high performance, thanks to the specially designed inter-pixel relation module [47], graph partition algorithm [50], salient detector [49], fully-supervised model retraining [46, 47]. Unlike previous methods [50, 47, 42], JTSM is end-to-end trainable. When ResNet18-WS is used as backbone, JTSM achieves comparable performance with previous state-of-the-art methods.

4.4. Weakly Supervised Object Detection

We evaluate the detection performance of the proposed JTSM on all three datasets, in which we only used image-level labels of *thing* categories. We also remove the segmentation refinement branch, as panoptic mining branch already outputs detection results. Comparisons with recent state-of-the-art methods are listed in Tab. 8. With ResNet18-WS backbone, JTSM reaches the state-of-the-art mAP of 53.4% and 51.5% on VOC 2007 and VOC 2012. JTSM produces 9.4% mAP and 21.3% $m\text{AP}_{0.5}$ on MS-COCO.

Although JTSM is not specially designed for object detection, it shows surprising results and achieves state-of-the-art performance for many metrics. We attribute the performance gains to MoIPool, which enables to extract pixel-accurate feature maps of arbitrary-shape regions.

5. Conclusion

In this paper, we propose a Joint *Thing*-and-*Stuff* Ming (JTSM) framework to learn panoptic segmentation with only image-level labels for the first time. To achieve this goal, a novel mask of interest pooling (MoIPool) is proposed to extract pixel-accurate feature maps of arbitrary-shape regions, which outputs fix-size feature maps for all semantic categories with the same representational power. We further integrate the mined masks with bottom-up object evidence to improve spatial coherence and contour localization. Finally, additional instance and semantic segmentation are learned via self-train to refine panoptic segmentation. Experimental results on PASCAL VOC and MS COCO demonstrate the effectiveness of JTSM compared to strong baselines. As a by-product, JTSM achieves competitive results for weakly supervised object detection and instance segmentation.

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) and Beijing Nova Program under Grant Z201100006820023.

References

[1] 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

[2] Qizhu Li, Anurag Arnab, and Philip H. S. Torr. Weakly- and Semi-Supervised Panoptic Segmentation. In *European Conference on Computer Vision (ECCV)*, 2018. 1, 2, 3, 7

[3] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic Segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 6, 7

[4] Yanwei Li, Xinze Chen, Zheng Zhu, Lingxi Xie, Guan Huang, Dalong Du, and Xingang Wang. Attention-Guided Unified Network for Panoptic Segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[5] Yangxin Wu, Gengwei Zhang, Yiming Gao, Xiajun Deng, Ke Gong, Xiaodan Liang, and Liang Lin. Bidirectional Graph Reasoning Network for Panoptic Segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[6] 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, 6

[7] 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, 6

[8] Hakan Bilen and Andrea Vedaldi. Weakly Supervised Deep Detection Networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2, 4, 8

[9] 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. 2

[10] 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. 2, 8

[11] 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. 2, 8

[12] Eu Wern Teh and Yang Wang. Attention Networks for Weakly Supervised Object Localization. In *The British Machine Vision Conference (BMVC)*, 2016. 2

[13] 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. 2

[14] 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. 2, 8

[15] 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. 2, 8

[16] 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. 2

[17] 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. 2

[18] 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. 2, 4, 8

[19] 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. 2, 8

[20] 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. 2

[21] 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. 2, 8

[22] 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. 2

[23] 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. 2

[24] 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. 2, 8

[25] 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. 2

[26] 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. 2, 8

[27] 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. 2

[28] 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. 2, 8

[29] 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. 2, 8

[30] 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. 2, 8

[31] Xiaopeng Zhang, Yang Yang, and Jiashi Feng. Learning to Localize Objects with Noisy Labeled Instances. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2019. 2

[32] 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. 2

[33] Luis Felipe Zeni and Claudio Jung. Distilling Knowledge from Refinement in Multiple Instance Detection Networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2020. 2

[34] 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. 2

[35] Ke Yang and Zhiyuan Wang. Objectness Consistent Representation for Weakly Supervised Object Detection. In *ACMMM*, 2020. 2

[36] 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. 2

[37] 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. 2

[38] 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. 2

[39] 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. 2

[40] 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. 2

[41] 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. 3

[42] 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. 3, 7, 8

[43] 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. 3

[44] 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. 3, 7

[45] 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. 3, 7

[46] 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. 3, 7, 8

[47] 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. 3, 7, 8

[48] 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. 3

[49] 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. 3, 8

[50] 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. 3, 7, 8

[51] 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. 3, 7

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

[53] Deepak Pathak, Philipp Krähenbühl, and Trevor Darrell. Constrained Convolutional Neural Networks for Weakly Supervised Segmentation. In *IEEE International Conference on Computer Vision (ICCV)*, 2015. 3

[54] Wataru Shimoda and Keiji Yanai B. Distinct Class-Specific Saliency Maps for Weakly Supervised Semantic Segmentation. In *European Conference on Computer Vision (ECCV)*, 2016. 3

[55] Fatemehsadat Saleh, Ali Akbarian, Mohammad, Sadegh, Mathieu Salzmann, Lars Petersson, Stephen Gould, and Jose M. Alvarez. Built-in Foreground/Background Prior for Weakly-Supervised Semantic Segmentation. In *European Conference on Computer Vision (ECCV)*, 2016. 3

[56] Alexander Kolesnikov and Christoph H. Lampert. Seed, Expand and Constrain: Three Principles for Weakly-Supervised Image Segmentation. In *European Conference on Computer Vision (ECCV)*, 2016. 3

[57] Yunchao Wei, Huaxin Xiao, Honghui Shi, Zequn Jie, Jiashi Feng, and Thomas S. Huang. Revisiting Dilated Convolution: A Simple Approach for Weakly- and Semi- Supervised Semantic Segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[58] 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, 7

[59] Zilong Huang, Xinggang Wang, Jiasi Wang, Wenyu Liu, and Jingdong Wang. Weakly-Supervised Semantic Segmentation Network with Deep Seeded Region Growing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[60] Pedro O. Pinheiro and Ronan Collobert. From Image-level to Pixel-level Labeling with Convolutional Networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 3

[61] Arslan Chaudhry, Puneet K. Dokania, and Philip H. S. Torr. Discovering Class-Specific Pixels for Weakly-Supervised Semantic Segmentation. In *The British Machine Vision Conference (BMVC)*, 2017. 3

[62] Yu Zeng, Yunzhi Zhuge, Huchuan Lu, and Lihe Zhang. Joint Learning of Saliency Detection and Weakly Supervised Semantic Segmentation. In *IEEE International Conference on Computer Vision (ICCV)*, 2019. 3

[63] Yude Wang, Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Self-supervised Equivariant Attention Mechanism for Weakly Supervised Semantic Segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[64] Meng Tang, Abdelaziz Djelouah, Federico Perazzi, Yuri Boykov, and Christopher Schroers. Normalized Cut Loss for Weakly-supervised CNN Segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[65] Yuxing Tang, Josiah Wang, Xiaofang Wang, Boyang Gao, Emmanuel Dellandrea, Robert Gaizauskas, and Liming Chen. Visual and Semantic Knowledge Transfer for Large Scale Semi-Supervised Object Detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3

[66] Olga Veksler. Regularized Loss for Weakly Supervised Single Class Semantic Segmentation. In *European Conference on Computer Vision (ECCV)*, 2020. 3

[67] Xiaojuan Qi, Zhengzhe Liu, Jianping Shi, Hengshuang Zhao, and Jiaya Jia. Augmented feedback in semantic segmentation under image level supervision. In *European Conference on Computer Vision (ECCV)*, 2016. 3

[68] Yu-Ting Chang, Qiaosong Wang, Wei-Chih Hung, Robinson Piramuthu, Yi-Hsuan Tsai, and Ming-Hsuan Yang. Weakly-Supervised Semantic Segmentation via Sub-category Exploration. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[69] Nikita Araslanov and Stefan Roth. Single-Stage Semantic Segmentation from Image Labels. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[70] Jungbeom Lee, Eunji Kim, Sungmin Lee, Jangho Lee, and Sungroh Yoon. FickleNet: Weakly and Semi-supervised Semantic Image Segmentation using Stochastic Inference. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[71] Guolei Sun, Wenguan Wang, Jifeng Dai, and Luc Van Gool. Mining Cross-Image Semantics for Weakly Supervised Semantic Segmentation. In *European Conference on Computer Vision (ECCV)*, 2020. 3

[72] Xiang Wang, Shaodi You, Xi Li, and Huimin Ma. Weakly-Supervised Semantic Segmentation by Iteratively Mining Common Object Features. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[73] Junsong Fan, Zhaoxiang Zhang, Chunfeng Song, and Tieniu Tan. Learning Integral Objects with Intra-Class Discriminator for Weakly-Supervised Semantic Segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[74] Wataru Shimoda and Keiji Yanai. Self-Supervised Difference Detection for Weakly-Supervised Semantic Segmentation. In *IEEE International Conference on Computer Vision (ICCV)*, 2019. 3

[75] Paul Vernaza and Manmohan Chandraker. Learning Random-Walk Label Propagation for Weakly-Supervised Semantic Segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 3

[76] Liyi Chen, Weiwei Wu, Chenchen Fu, Xiao Han, and Yuntao Zhang. Weakly Supervised Semantic Segmentation with Boundary Exploration. In *European Conference on Computer Vision (ECCV)*, 2020. 3

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

[78] K. E. A. van de Sande, J. R. R. Uijlings, T. Gevers, and A. W. M. Smeulders. Segmentation As Selective Search for Object Recognition. In *IEEE International Conference on Computer Vision (ICCV)*, 2011. 4

[79] 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

[80] Ross Girshick. Fast R-CNN. In *IEEE International Conference on Computer Vision (ICCV)*, 2015. 4, 7

[81] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *IEEE International Conference on Computer Vision (ICCV)*, 2017. 4, 5, 7, 8

[82] Tien-Ju Yang, Maxwell D. Collins, Yukun Zhu, Jyh-Jing Hwang, Ting Liu, Xiao Zhang, Vivienne Sze, George Papandreou, and Liang-Chieh Chen. DeeperLab: Single-Shot Image Parser. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 6, 7

[83] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic Feature Pyramid Networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 7

[84] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2015. 8

[85] Zhaoyang Zeng, Bei Liu, Jianlong Fu, Hongyang Chao, and Lei Zhang. WSOD^2: Learning Bottom-up and Top-down Objectness Distillation for Weakly-supervised Object Detection. In *IEEE International Conference on Computer Vision (ICCV)*, 2019. 8

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

[87] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid Scene Parsing Network. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 6

[88] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic Contours from Inverse Detectors. In *IEEE International Conference on Computer Vision (ICCV)*, 2011. 6

[89] 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

[90] 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. 6