Beyond Semantic to Instance Segmentation: Weakly-Supervised Instance Segmentation via Semantic Knowledge Transfer and Self-Refinement

Beomyoung Kim[1] Youngjoon Yoo[1,2] Chae Eun Rhee[3] Junmo Kim[4]

NAVER CLOVA[1] NAVER AI Lab[2] Inha University[3] KAIST[4]

Abstract

Weakly-supervised instance segmentation (WSIS) has been considered as a more challenging task than weakly-supervised semantic segmentation (WSSS). Compared to WSSS, WSIS requires instance-wise localization, which is difficult to extract from image-level labels. To tackle the problem, most WSIS approaches use off-the-shelf proposal techniques that require pre-training with instance or object level labels, deviating the fundamental definition of the fully-image-level supervised setting. In this paper, we propose a novel approach including two innovative components. First, we propose a semantic knowledge transfer to obtain pseudo instance labels by transferring the knowledge of WSSS to WSIS while eliminating the need for the off-the-shelf proposals. Second, we propose a self-refinement method to refine the pseudo instance labels in a self-supervised scheme and to use the refined labels for training in an online manner. Here, we discover an erroneous phenomenon, semantic drift, that occurred by the missing instances in pseudo instance labels categorized as background class. This semantic drift occurs confusion between background and instance in training and consequently degrades the segmentation performance. We term this problem as semantic drift problem and show that our proposed self-refinement method eliminates the semantic drift problem. The extensive experiments on PASCAL VOC 2012 and MS COCO demonstrate the effectiveness of our approach, and we achieve a considerable performance without off-the-shelf proposal techniques. The code is available at https://github.com/clovaai/BESTIE.

1. Introduction

The recent line of weakly-supervised semantic segmentation (WSSS) approaches [21,28,29,45] have achieved impressive performance enhancement, with commonly using class activation maps (CAMs) [46] to obtain class-wise localization maps from image-level labels. However, weakly-supervised instance segmentation (WSIS) using image-level labels is still an open task because the CAM does not provide instance-wise localization maps.

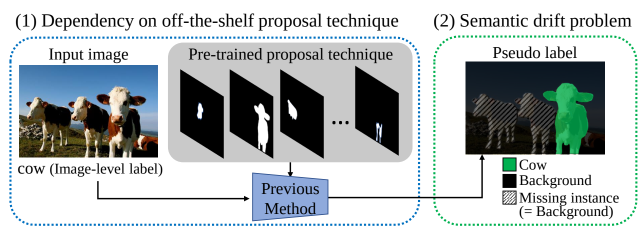

Figure 1. Two limitations of previous WSIS methods: (1) dependency on off-the-shelf proposal technique that requires pre-training with high-level labels including object or instance information. (2) semantic drift problem caused by the missing instances in the pseudo label guided as *background* class. Our proposed BESTIE aims to solve the two problems simultaneously.

To extract the instance-wise information, most WSIS methods use off-the-shelf proposal techniques. PRM [47] takes suitable instance masks from segment proposals generated by MCG [38] and generates pseudo instance labels. Also, LIID [33] utilizes a pre-trained salient instance segmentor [12] which produces class-agnostic instance-level masks. We note that MCG and salient instance segmentor used in each method require pre-training with object boundary labels and class-agnostic instance masks, respectively.

However, the proposal-guided WSIS methods have two limitations as shown in Figure 1. First, their dependency on the off-the-shelf proposal techniques is considerably high, and it makes the methods difficult to apply to other specific domains such as medical images since the proposal techniques mostly target general objects. Furthermore, in a strict sense, the use of the proposal techniques trained by object or instance level information deviates from the definition of fully-image-level supervised segmentation. Second, these methods cannot cope with the performance degradation caused by noisy pseudo-labels containing missing instances (*i.e.*, false-negatives). As shown in Figure 1, the left two missing cows are guided to the *background* class and the right cow is guided to the *cow* class, although all cows have semantically similar visual cues. We call this problem as *semantic drift problem*. This semantic drift between background and instance confuses the network and deteriorates the stable training convergence.

In this paper, we propose a new WSIS method, **BESTIE**: BEyond Semantic segmentation To InstancE Segmentation. BESTIE deprecates the use of off-the-shelf techniques to strictly follow a fully-image-level supervised setting. Also, BESTIE alleviates the semantic drift problem. To solve the two problems, BESTIE proposes two innovative components, *semantic knowledge transfer* and *self-refinement*.

Specifically, in *semantic knowledge transfer*, we transfer the knowledge of WSSS, which is relatively profoundly studied, to WSIS to generate the rough pseudo instance labels. To obtain the instance cues from image-level labels, we propose *peak attention module (PAM)* that makes the CAM highlight the sparse representative region of objects. We note that the proposed components only use image-level labels, including the WSSS, and this eliminates the need for the off-the-shelf proposal technique. Furthermore, to address the semantic drift problem, we introduce *instance-aware guidance* that dynamically assigns the guidance region only to the labeled instance region. This strategy allows more stable training of the network and progressively captures instance-level information of the missing instances. Along with this strategy, to further refine the pseudo labels, we propose the self-supervised instance label refinement method that converts false-negatives in the pseudo labels to true-positives by a self-supervised manner and reflects them to the training in an online manner. This method, shortly named *self-refinement*, improves the quality of the pseudo labels as training progresses.

The extensive experiments on PASCAL VOC 2012 [11] and MS COCO 2017 [31] show the effectiveness of the proposed design. Even without the off-the-shelf proposal techniques, our method achieves a state-of-the-art performance of 51.0% mAP_{50} on VOC 2012 and 28.0% AP_{50} on COCO dataset. Furthermore, we model the point-supervised instance segmentation by replacing the instance cues with point labels and can further boost the performance with an economical annotation cost.

Our contribution can be summarized as follows:

- We propose a novel WSIS method only using image-level labels, strictly following the fully-image-level supervised setting without the help of the proposal techniques pre-trained by object or instance level labels.
- We design the *semantic knowledge transfer* strategy to obtain pseudo instance labels. This transfers the knowledge of WSSS and instance cues, extracted from the proposed PAM, to WSIS while eliminating the use of off-the-shelf proposal techniques.
- We propose a *self-refinement* method to refine the pseudo instance labels in a self-supervised manner and reflect them back to the training in an online manner. Here, we introduce the *instance-aware guidance* strategy to resolve the *semantic drift* problem newly discovered in this paper.

2. Related Work

2.1. Weakly-Supervised Semantic Segmentation

Most weakly-supervised semantic segmentation (WSSS) studies handling image-level labels use CAMs [46] to localize class-wise object regions. However, CAMs mainly focus on the sparse and discriminative object regions. To address this issue, recent WSSS methods have proposed many approaches to expand activation regions. AE-PSL [44] removes discriminative object regions by shifting attention to adjacent non-discriminative regions. DRS [21] proposes a module suppressing discriminative regions to expand activation regions. However, these approaches are dependent on the off-the-shelf guidance, $i.e.$, saliency map. To eliminate the use of the saliency map, saliency map-free methods has been also proposed: RRM [45] proposes an end-to-end network jointly producing both CAMs and segmentation output to generate pseudo labels from only reliable pixels. PMM [29] proposes the proportional pseudo-mask generation by variation smoothing. Also, some approaches have shown the expansion possibility of the WSSS to various domains such as medical [6] and satellite [35] images.

2.2. Instance Segmentation

Unlike semantic segmentation categorizing the pixel region in class-level, instance segmentation requires an instance-level mask. The most widely used approach is a box-based two-stage method, $e.g.$, Mask R-CNN [15], that predicts bounding boxes and then extracts the instance mask for each bounding box; this approach has reigned on the throne with state-of-the-art performance. Recently, for the simple instance segmentation process, box-free one-stage instance segmentation methods [7, 34] have been proposed. They represent each instance using 2-dimensional (2D) offset vectors; the pixels covering each instance are represented as 2D offset vectors directed to the center of each instance. The center point of each instance is extracted from the center heatmap [7] or by clustering 2D offset vectors [34] and the instance mask is obtained from instance grouping with the center point and 2D offset vectors.

2.3. Weakly-Supervised Instance Segmentation

The major difficulty in solving weakly-supervised instance segmentation (WSIS) occurs in the process of obtaining instance-level information from image-level labels. To solve the problem, PRM [47] produces a peak response map

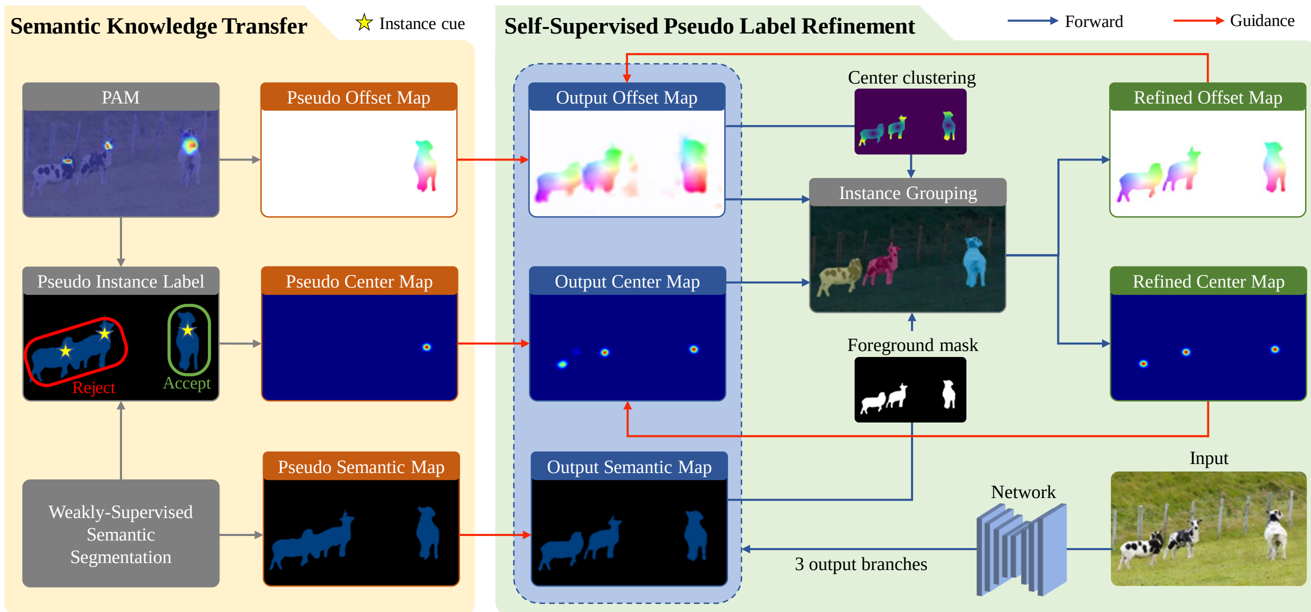

Figure 2. The overview of our framework including two-innovated steps: *semantic knowledge transfer* and *self-refinement*. In *semantic knowledge transfer*, we obtain pseudo instance labels from the knowledge of WSSS and instance cues. Here, we obtain instance cues using *peak attention module (PAM)*. In the *self-refinement*, the network refines the pseudo labels in a self-supervised scheme and reflects them to the training in an online manner. For the stable training alleviating the semantic drift problem, we apply *instance-aware guidance* strategy. We note that the entire process only uses the image-level labels.

using the proposed peak back-propagation and then selects appropriate segment proposals generated by MCG [38]. Arun *et al*. [3] defines the uncertainty in the pseudo label generation with the help of the segment proposals and iteratively train the network using the pseudo labels in an offline manner. Fan *et al*. [13] and LIID [33]) use the salient instance segmentor [12] that produces class-agnostic instance-level masks when they generate pseudo labels. In addition, box-based two-stage approaches [14, 19] use a selective search [42] method to generate box proposals. However, the off-the-shelf proposal techniques require pre-training with high-level supervision: class-agnostic object boundary for MCG and class-agnostic instance mask for the salient instance segmentor. Also, since these proposal techniques target general objects, it disturbs their utilization in other domains such as a medical image. IRN [1] proposes a proposal-free method focusing on class-equivalence relations between a pair of pixels and represents instance-level information using their displacement field. However, IRN has trouble in obtaining accurate instance-level information because the inter-pixel relations used in IRN are based on inter-class, not inter-instance. To the best of our knowledge, existing methods have not considered the semantic drift problem caused by missing instances in pseudo labels, one fundamental challenge that should be addressed for WSIS. In this paper, we discover and tackle the semantic drift problem explicitly for the first time and achieve an improved result in a fully-image-level supervised setting.

3. Proposed Method

3.1. Overview

As shown in the left part of Figure 2, we first obtain pseudo instance labels using the knowledge of WSSS and instance cues, and this process is called *semantic knowledge transfer*. Here, we extract instance cues from image-level labels using the proposed *peak attention module (PAM)* module. Then, we apply the self-supervised pseudo label refinement abbreviated as *self-refinement* strategy that refines the pseudo instance labels in a self-supervised scheme and reflects them to the training in an online manner as described in the right part of Figure 2. In order to resolve the semantic drift problem and ensure stable training, we introduce the *instance-aware guidance* strategy. We note that our framework only uses the image-level labels as our guidance source including the WSSS part to deprecate off-the-shelf proposal techniques in the entire process. We also provide the Pytorch-style pseudo-code for each proposed component in our supplementary material, showing that our method is quite easy to be implemented and simple.

3.2. Preliminary: Instance Representation

Motivated by Panoptic-DeepLab [7], we represent an instance as a center point and corresponding 2D offset vectors. The 2D offset vectors direct the center point of each instance. By adopting this representation method, we con-

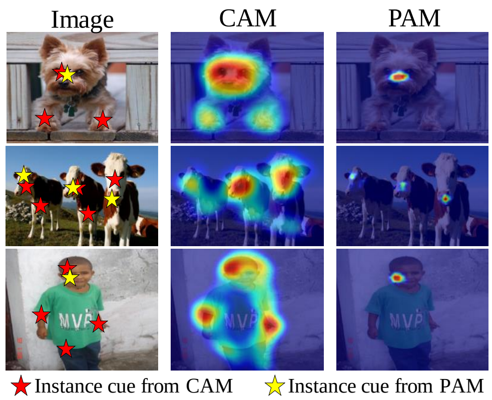

Figure 3. The activation maps from CAM and PAM. PAM helps extract more accurate instance cues than CAM.

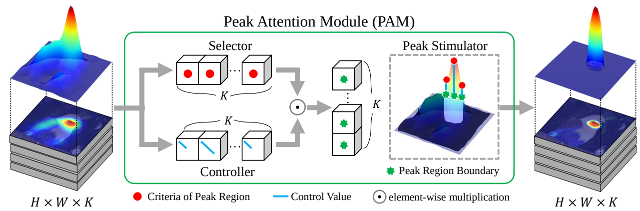

Figure 4. PAM architecture. From the selector, criteria of peak regions are selected. For a better explanation, three of them are illustrated in red points. Then, the controller determines how much to strengthen the attention on peak regions with control values, and each value is illustrated as the length of a blue line. Using criteria points and control values, the boundary of peak regions is set, and the stimulator strengthens the attention on peak regions by deactivating noisy regions whose values are lower than the boundary.

struct an instance segmentation network following the architecture of Panoptic-DeepLab, which consists of three output branches: semantic segmentation map, center map, and offset map. The semantic segmentation map determines the foreground region. For the post-processing, we extract center points of each instance from the center map; pixel locations with the same value before and after the max-pooling of the center map are regarded as center points. Then, we allocate the ID of each instance in a pixel-level, and this procedure is called as instance grouping; when the extracted n-th center point is denoted as (x_n, y_n) and the offset map is denoted as $\mathcal{O}(i,j)$ at the pixel location (i,j), the instance ID $k_{i,j}$ at the pixel (i,j) becomes

$$k_{i,j} = \underset{k}{\mathrm{argmin}} ||(x_k, y_k) - ((i,j) + \mathcal{O}(i,j))||. \quad (1)$$

3.3. Semantic Knowledge Transfer

The proposed *semantic knowledge transfer* transfers the knowledge of WSSS to WSIS. For the transfer, we rethink the two conditions of semantic and instance segmentation. The **first condition** is that instance segmentation should separate overlapping instances of the same class, unlike semantic segmentation. The **second condition** is that semantic segmentation and instance segmentation are equivalent to each other when instances of the same class do not overlap. Based on these conditions, we generate pseudo instance labels utilizing the knowledge of WSSS. From the WSSS output and instance presence cues, we check whether instances overlap. Then, we select a non-overlapping instance mask as a pseudo instance mask, described in the left part of Figure 2. Specifically, by performing the connected component labeling (CCL) [16] algorithm on the WSSS output of each class, we obtain instance mask candidates and check how many instance cues are included in each instance mask candidate. Following the **second condition**, the instance mask candidate with only one instance cue is selected as the pseudo instance mask.

For the proper *semantic knowledge transfer*, we require the accurate instance cue extraction method only using image-level labels. The previous work, PRM [47], extracts the instance cue from CAMs [46]. However, CAMs have a limitation in obtaining the accurate instance cue because several instance cues might be extracted in a single instance due to noisy activation regions as illustrated in Figure 3. It disturbs the generation of pseudo instance labels since it violates the **second condition**.

To address this limitation, we propose a *peak attention module (PAM)* to extract one appropriate instance cue per instance, motivated by DRS [21]. DRS suppresses discriminative object regions, spreading attention to adjacent non-discriminative regions in a self-supervised manner. Contrary to the DRS, our PAM aims to strengthen the attention on peak regions, while weakening the attention on noisy activation regions. PAM consists of three parts, as illustrated in Figure 4: *selector*, *controller*, and peak *stimulator*. We denote the intermediate feature map as $X \in \mathbb{R}^{H \times W \times K}$, where H, W, and K are the height, width, and the number of channels of X, respectively. The *selector* selects criteria points of peak regions using the global max pooling of X, and the criteria points are denoted as $S_p \in \mathbb{R}^{1 \times 1 \times K}$. The *controller* determines how much strengthen the attention on peak regions and is denoted as $G_p \in [0,1]^{1 \times 1 \times K}$. We strengthen the attention on peak regions by deactivating the attention on noisy regions. In particular, $\tau_p = S_p \cdot G_p$ plays a role of the boundary of peak regions where $\cdot$ is an element-wise multiplication. The regions in X higher than τ_p are regarded as peak regions, otherwise regarded as noisy regions. We deactivate the noisy regions by setting the value to zero, focusing on peak regions. For the controller, we adopt the non-learnable setting of DRS, that is, all elements of G_p are set to a constant value α; the α is set to 0.7, and we find out that changes in the α between 0.3 and 0.7 do not significantly affect the performance of WSIS. The PAM is plugged into the classifier, and we produce activation maps

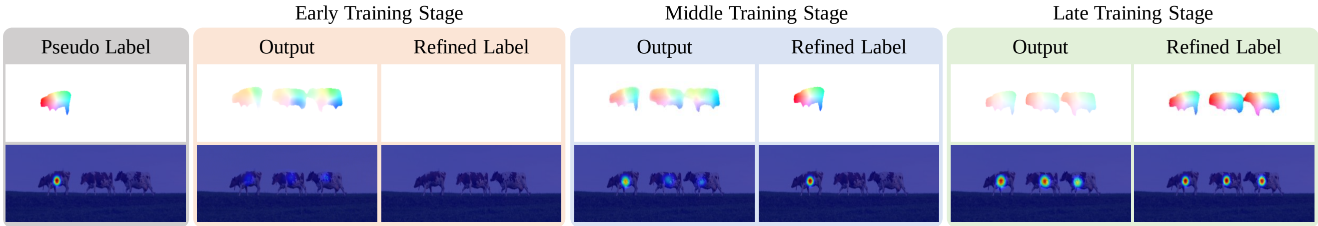

Figure 5. Comparison of offset and center maps. As training iteration progresses, the network generates the higher-quality refined label.

that localize the sparse representative region of each object as shown in Figure 3 and then extract local maximum points ($i.e.$, instance cues). Note that our PAM does not require additional training parameters, and the classifier with PAM is optimized with a binary cross-entropy objective function by adaptively focusing on peak regions while increasing classification ability.

Combined with the knowledge of WSSS and instance cues extracted from PAM, we obtain pseudo instance masks and convert these masks into the pseudo center and offset maps following our instance representation method, as illustrated in Figure 2. For the center map, the centroid point of each pseudo instance mask is encoded in a 2D Gaussian kernel with a standard deviation of 6 pixels. For the offset map, all pixels in the pseudo instance mask contain 2D offset vectors directed to the corresponding center point.

3.4. Instance-aware Guidance

When training with the pseudo instance labels obtained by the *semantic knowledge transfer*, we should handle the semantic drift problem. Since the missing instances in the pseudo labels are guided as a *background* class, semantic drift between background and instance deteriorates the stable training convergence. To alleviate this problem, we introduce an *instance-aware guidance* by taking the advantage of our instance representation method.

In our instance representation in Section 3.2, the offset and center maps represent instance-level information within the foreground region determined by the semantic segmentation map. It means that the background region of the offset and center maps can be regarded as ignored regions. Correspondingly, we dynamically assign the guidance region for the offset and center maps to only the region for the labeled instances; this strategy is called the *instance-aware guidance* and helps alleviate the semantic drift problem because the region of the offset and center maps for the missing instances ($e.g.$, the white region of the pseudo offset map in Figure 2) is not reflected in the objective function. Consequently, as shown in Figure 5, the network can stably capture the instance-level information of the missing instances as the training progresses.

3.5. Self-Supervised Pseudo Label Refinement

Even we can alleviate the semantic drift problem, the number of true-positives in the pseudo labels is still not enough for training the network. For example, we can only have 30% true-positives in the pseudo labels for VOC 2012. Here, we propose a self-supervised pseudo label refinement strategy, abbreviated as *self-refinement*, that refines the pseudo labels by converting false-negatives to true-positives in a self-supervised manner and reflects the refined labels to the training in an online manner. The overall process of the *self-refinement* is illustrated on the right side of Figure 2. First, by training with the pseudo instance labels using *instance-aware guidance* strategy, the network stably develops the generalization ability and gradually captures the instance-level information of the missing instances ($i.e.$, false-negatives). Next, following the Eq. (1), we perform the instance grouping using the network outputs. Then, we generate refined offset and center maps from the instance mask created by instance grouping. Last, the refined maps are used as guidance for the network.

For better refinement, we extract the center point by clustering the 2D offset vectors in the output offset map. We call this process a *center clustering* and explain the detailed algorithm in the supplementary material. Even if the output center map misses some center points, we complement the refined center map using the clustered center points.

Using both pseudo labels and refined labels, we train the network. We denote the network output semantic segmentation, offset map, and center map as $\mathcal{S}(\cdot)$, $\mathcal{O}(\cdot)$, and $\mathcal{C}(\cdot)$, respectively. For the *instance-aware guidance* of the offset and center maps, we collect sets of pixels for labeled instance regions from pseudo and refined labels, and each set is denoted as $\mathcal{P}_{pseudo}$ and $\mathcal{P}_{refined}$. To utilize the refined labels as soft labels, we design a weight mask $\mathcal{W}(i,j)$:

$$\mathcal{W}^n(i,j) = \begin{cases} \mathcal{C}(x_n, y_n) & (i,j) \in \mathcal{P}^n_{pseudo}, \\ 0 & \text{otherwise}, \end{cases} \tag{2}$$

where the center point of n-th instance in the refined labels are denoted as (x_n, y_n), and $\mathcal{C}(x_n, y_n)$ means the confidence score of the n-th instance. The $\mathcal{W}$ is used as the weight of the objective function for the refined labels. The objective function of the center map is defined as:

$$\mathcal{L}_{center} = \frac{1}{|\mathcal{P}_{pseudo}|} \sum_{(i,j)\in\mathcal{P}_{pseudo}} (\mathcal{C}(i,j) - \hat{\mathcal{C}}(i,j))^2 + \frac{1}{|\mathcal{P}_{refined}|} \sum_{(i,j)\in\mathcal{P}_{refined}} \mathcal{W}(i,j)\cdot(\mathcal{C}(i,j) - \bar{\mathcal{C}}(i,j))^2, \tag{3}$$

where the pseudo and refined center maps are $\hat{\mathcal{C}}(i,j)$ and $\bar{\mathcal{C}}(i,j)$, respectively. Also, the objective function of the offset map is defined as:

$$\mathcal{L}_{offset} = \frac{1}{|\mathcal{P}_{pseudo}|} \sum_{(i,j)\in\mathcal{P}_{pseudo}} |\mathcal{O}(i,j) - \hat{\mathcal{O}}(i,j)| + \frac{1}{|\mathcal{P}_{refined}|} \sum_{(i,j)\in\mathcal{P}_{refined}} \mathcal{W}(i,j)\cdot|\mathcal{O}(i,j) - \bar{\mathcal{O}}(i,j)|, \tag{4}$$

where the pseudo and refined offset maps are $\hat{\mathcal{O}}(i,j)$ and $\bar{\mathcal{O}}(i,j)$, respectively. Lastly, the objective function of the segmentation map is defined as:

$$\mathcal{L}_{sem} = -\frac{1}{|\mathcal{P}_{sem}|} \sum_{(i,j)\in\mathcal{P}_{sem}} log\mathcal{S}(i,j), \tag{5}$$

where $\mathcal{S}$ is the output semantic map and $\mathcal{P}_{sem}$ is the set of all pixels in $\mathcal{S}$. The network is jointly trained with the above objective functions, and the final objective function is:

$$\mathcal{L} = \lambda_{center}\mathcal{L}_{center} + \lambda_{offset}\mathcal{L}_{offset} + \lambda_{sem}\mathcal{L}_{sem}, \tag{6}$$

where λ is a weight parameter, and set $\lambda_{center} = 200$, $\lambda_{offset} = 0.01$, and $\lambda_{sem} = 1$ as used in [7].

Through the *self-refinement* strategy, the pseudo labels can convert into high-quality refined labels. The refined labels are generated from the network in an online manner at every mini-batch. Since most of the operations of *self-refinement* can be performed on GPU, the throughput of the *self-refinement* is small.

4. Experiments

4.1. Dataset and Evaluation Metrics

We demonstrate the effectiveness of the proposed approach on Pascal VOC 2012 [11] and COCO [31] datasets. For VOC 2012 dataset, following the common practice in previous works [1, 33], we use the augmented dataset that contains 10,582 training and 1,449 validation images with 20 object categories. COCO dataset consists of 115K training, 5K validation, and 20K testing images with 80 object categories. We evaluate the performance using the mean average precision (mAP) with intersection-over-union (IoU) thresholds of 0.25, 0.5, 0.7, and 0.75 for VOC 2012 and averaged AP over IoU thresholds from 0.5 to 0.95 for COCO.

Table 1. Comparison of state-of-the-art WSIS methods on VOC 2012 *val*-set. † indicates applying MRCNN refinement. We denote the supervision sources as: $\mathcal{F}$ (full mask), $\mathcal{I}$ (image-level label), $\mathcal{P}$ (point), $\mathcal{C}$ (object count). The off-the-shelf proposal techniques are denoted as follows: $\mathcal{M}$ (segment proposal [38]), $\mathcal{R}$ (region proposal [42]), $\mathcal{S_I}$ (salient instance segmentor [12]).

Method	Sup	Extra	mAP_{25}	mAP_{50}	mAP_{70}	mAP_{75}
Mask R-CNN [15]	$\mathcal{F}$	-	76.7	67.9	52.5	44.9
PRM [47]	$\mathcal{I}$	$\mathcal{M}$	44.3	26.8	-	9.0
IAM [48]	$\mathcal{I}$	$\mathcal{M}$	45.9	28.3	-	11.9
Label-PEnet [14]	$\mathcal{I}$	$\mathcal{R}$	49.2	30.2	-	12.9
CL [19]	$\mathcal{I}$	$\mathcal{M}, \mathcal{R}$	**56.6**	38.1	-	12.3
BESTIE (ours)	$\mathcal{I}$	-	53.5	**41.8**	**28.3**	**24.2**
OCIS [9]	$\mathcal{C}$	$\mathcal{M}$	48.5	30.2	-	14.4
WISE-Net [23]	$\mathcal{P}$	$\mathcal{M}$	53.5	43.0	-	25.9
BESTIE (ours)	$\mathcal{P}$	-	**58.6**	**46.7**	**33.1**	**26.3**
WISE† [24]	$\mathcal{I}$	$\mathcal{M}$	49.2	41.7	-	23.7
IRN† [1]	$\mathcal{I}$	-	-	46.7	23.5	-
LIID† [33]	$\mathcal{I}$	$\mathcal{M}, \mathcal{S_I}$	-	48.4	-	24.9
Arun *et al.*† [3]	$\mathcal{I}$	$\mathcal{M}$	59.7	50.9	30.2	**28.5**
BESTIE (ours)†	$\mathcal{I}$	-	**61.2**	**51.0**	**31.9**	26.6
BESTIE (ours)†	$\mathcal{P}$	-	**66.4**	**56.1**	**36.5**	**30.2**

Table 2. Quantitative comparison of state-of-the-art WSIS methods on MS COCO 2017 dataset.

Method	Sup	Extra	AP	AP_{50}	AP_{75}
COCO val2017					
Mask R-CNN [15]	$\mathcal{F}$	-	35.4	57.3	37.5
WS-JDS [39]	$\mathcal{I}$	$\mathcal{M}$	6.1	11.7	5.5
WISE-Net [23]	$\mathcal{P}$	$\mathcal{M}$	7.8	18.2	8.8
BESTIE (ours)	$\mathcal{I}$	-	**14.3**	**28.0**	**13.2**
BESTIE (ours)	$\mathcal{P}$	-	**17.7**	**34.0**	**16.4**
COCO test-dev					
Mask R-CNN [15]	$\mathcal{F}$	-	35.7	58.0	37.8
Fan *et al.* [13]	$\mathcal{I}$	$\mathcal{S_I}$	13.7	25.5	13.5
LIID [33]	$\mathcal{I}$	$\mathcal{M}, \mathcal{S_I}$	**16.0**	27.1	**16.5**
BESTIE (ours)	$\mathcal{I}$	-	14.4	**28.0**	13.5
BESTIE (ours)	$\mathcal{P}$	-	**17.8**	**34.1**	**16.7**

4.2. Implementation Details

For the *semantic knowledge transfer*, we extract instance cues from the classifier equipped with our PAM. We describe the detailed architecture of the classifier and analysis in the supplementary material. For the fully-image-level supervised setting, we adopt PMM [29] as our WSSS method because it does not use the saliency maps.

For the instance segmentation network, we follow the network structure of Panoptic-DeepLab [7] with a modification. We change the center map from class-agnostic to class-wise for more accurate instance grouping. We adopt HRNet48 [40] as our backbone network. The input size for training is 416×416, and we keep the original resolution for evaluation. We train the network for 70 epochs with 32 batch size using Adam optimizer [22] with 5×10^{-5} learning rate and polynomial learning rate scheduling [32]. Some approaches [1, 33] employ an additional training step on Mask R-CNN [15]; we denote this step as MRCNN refinement and train the Mask R-CNN following the official

Table 3. Effect of the proposed methods: PAM, IAG (*instance-aware guidance*), refine (*self-refinement*), and cluster (*center clustering*).

PAM	IAG	refine	cluster	mAP_{50}
✗	✗	✗	✗	12.9
✓	✗	✗	✗	29.3
✓	✓	✗	✗	39.2
✓	✓	✓	✗	41.5
✓	✓	✓	✓	41.8

Table 4. Analysis of the effect of WSSS result on our WSIS performance. We measure the mAP_{50} instance segmentation performance according to WSSS methods.

Semantic Segmentation		Instance Segmentation
WSSS method	mIoU	mAP_{50}
SingleStage [2]	62.7	39.7
RRM [45]	66.3	41.1
PMM [29]	70.0	41.8
ground truth	-	49.4

Table 5. Effect of iterative training strategy on our method. *iters*=0 means without the iterative training strategy.

# iters	mAP_{50}
0	41.8
1	41.9
2	41.9

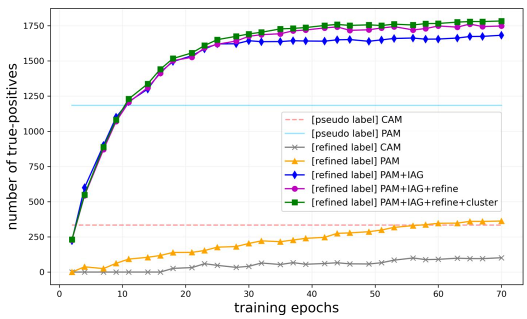

Figure 6. Evolution of the number of true-positives on VOC 2012 *train*-set.

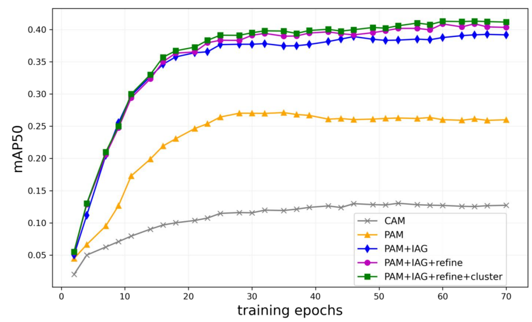

Figure 7. Evolution of the mAP_{50} on VOC 2012 *val*-set.

training recipe using pseudo labels generated by our network. We used PyTorch 1.7 framework [37] with CUDA 10.1, CuDNN 7 and 8 NVIDIA V100 GPUs.

4.3. Point-Supervised Instance Segmentation

In our framework, point label can be used as weak supervision. According to [4, 5], annotation costs are as follow: image-level (20.0 sec/img), object count (22.2 sec/img), point (23.3 sec/img), bounding box (38.1 sec/img), full mask (239.7 sec/img). The point label is an economical label that is 16% more expensive than the image-level label. For our point-supervised setting, we replace instance cues from PAM with point labels for the *semantic knowledge transfer* and replace the pseudo and refined center maps with the ground-truth center map for the *self-refinement*. Using the point supervision, we can obtain 10% more true-positives in pseudo labels for VOC 2012 due to the accurate instance cue, boosting the performance as in Table 1.

4.4. State-of-the-arts Comparison

We compare our BESTIE with existing state-of-the-art WSIS methods in Table 1 for VOC 2012 and Table 2 for COCO dataset. Even without the off-the-shelf proposals, BESTIE outperforms existing methods, especially in the AP_{50} metric. Although LIID [33] achieved a 1.6% AP higher than ours on COCO, they utilized two proposal techniques that require pre-training with high-level labels, violating the fully-image-level supervised setting. Compared to the fully-image-level supervised method, IRN [1], we outperform the method (51.0% $v.s.$ 46.7%) because their displacement field, which is similar to our offset map, does not consider the semantic drift problem. Also, IRN often fails to segment overlapping instances as in Figure 8 because their inter-pixel relations are derived from class-wise information, not instance-wise information. Given point supervision, we further increased the performance gap with other methods at a reasonable cost and achieve a new state-of-the-art performance on VOC and COCO datasets compared to the previous best point-supervised method, Wise-Net [24].

4.5. Ablation Study and Analysis

For analysis, we skip the Mask R-CNN refinement and follow the mentioned implementation details. We count the number of true-positives on VOC 2012 train set containing 1,464 images and 3,507 instances in Figure 6 and measure the mAP_{50} on VOC 2012 validation set in Figure 7.

Effect of PAM: As shown in Figure 3, we have trouble in obtaining accurate instance cues from conventional CAM due to the noisy activation regions. However, with our PAM, we can extract appropriate instance cues, which tremendously helps the proper *semantic knowledge transfer* and obtain three times more true-positive training samples than the CAM as shown in Figure 6, giving a 16.4% improvement as in the first and second rows of Table 3.

Effect of Instance-aware Guidance: In this section, we abbreviate *instance-aware guidance* as IAG. For analysis, we train the network without IAG; it means that the whole region (including the background region) of the offset and center maps are reflected in the objective function. As in the first and third rows of Table 3, without IAG, the performance drops by 9.9% because it suffers from the semantic

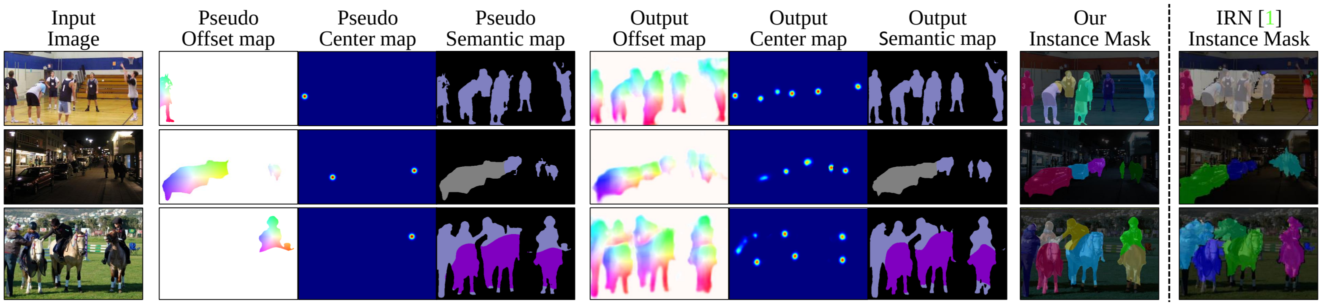

Figure 8. Qualitative results of our BESTIE trained with image-level supervision on VOC 2012 dataset.

drift problem as mentioned in the method section. Also, as shown in Figure 7, the model without IAG seems to be stuck in a local minimum, whereas the model with IAG seems to avoid the local minimum, enhancing the performance as training progresses. This result convinces us that IAG is effective to alleviate the semantic drift problem.

Effect of Self-Refinement: Here, we compare the result of the network trained using only the pseudo labels without the *self-refinement*. As shown in Figure 6, the network with the *self-refinement* can have more true-positives as training progresses. Since the refined labels from the *self-refinement* are guided to the training, the network can further capture instance-level features and improve the performance by 2.3% as in the third and fourth rows of Table 3.

Effect of Center Clustering: As in the fourth and last rows of Table 3, a 0.3% improvement when using the center clustering demonstrates that the center clustering can complement the generation of the refined labels.

Influence of WSSS method: The results in Table 4 shows how the WSSS result affects the WSIS. Originally, we adopt PMM [29] for our WSSS method, which shows 70.0% mIoU on VOC 2012 validation set. Adopting the WSSS methods of 7.3% and 3.7% lower mIoU (SingleStage [2] and RRM [45]) drops mAP_{50} by 2.1% and 0.7%. The results show that the performance of WSIS is robust to that of WSSS. Additionally, we train with ground-truth semantic segmentation labels and obtain a performance gain of 7.6% mAP_{50}; this result leaves us the opportunity that the advancement of the WSSS method can improve our method.

Does iterative training help? Some weakly-supervised methods [3, 20, 43] maximize their performance by the iterative training strategy; they generate pseudo labels when training is finished and re-train the network using the pseudo labels in an offline manner. This strategy gives a progressive improvement but requires a huge training complexity. However, this strategy does not give us a noticeable improvement as in Table 5, and we show that our single-step online *self-refinement* is quite efficient for label refinement.

Qualitative Results: We provide some qualitative results in Figure 8. Although the pseudo labels contain a few instance labels, BESTIE can accurately represent instance-level information, achieving high-quality instance masks.

Limitation and Future direction: Despite the significant performance enhancement from the proposed BESTIE, it remains more room to be improved. In our method, the number of true-positives in the pseudo labels is limited by overlapping objects in the image (see Figure 8). The number of overlapping objects is different from dataset to dataset, $i.e.$, less for VOC dataset but many for COCO dataset, and this affects the performance in some sort. Although our method achieved promising results in the VOC dataset with small 30% of true-positives in the pseudo labels, one future direction will be the suggestion of more effective true-positive acquiring rules for the various condition of data.

5. Conclusion

We proposed a novel approach for WSIS by addressing the pain-points of previous methods: dependency on the off-the-shelf proposals and the semantic drift problem. In our *semantic knowledge transfer*, we transferred the knowledge of WSSS combined with instance cues to WSIS and obtained pseudo instance labels. Here, we proposed the PAM to extract the instance cues. In our *self-refinement*, we refined the pseudo labels in a self-supervised scheme and employed them in training. To resolve the semantic drift problem, we introduced the *instance-aware guidance* strategy. Our approach outperforms the previous methods with only using image-level labels and without any off-the-shelf proposals. Lastly, we conclude that this research does not contain potential negative societal impact.

6. Acknowledgement

This work was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2021-0-02068, Artificial Intelligence Innovation Hub) and Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (NRF-2021R1A2C2008946).

References

[1] Jiwoon Ahn, Sunghyun Cho, and Suha Kwak. Weakly supervised learning of instance segmentation with inter-pixel relations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2209–2218, 2019. 3, 6, 7

[2] Nikita Araslanov and Stefan Roth. Single-stage semantic segmentation from image labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4253–4262, 2020. 7, 8

[3] Aditya Arun, CV Jawahar, and M Pawan Kumar. Weakly supervised instance segmentation by learning annotation consistent instances. In *European Conference on Computer Vision*, pages 254–270. Springer, 2020. 3, 6, 8

[4] Amy Bearman, Olga Russakovsky, Vittorio Ferrari, and Li Fei-Fei. What's the point: Semantic segmentation with point supervision. In *European conference on computer vision*, pages 549–565. Springer, 2016. 7

[5] Míriam Bellver Bueno, Amaia Salvador Aguilera, Jordi Torres Viñals, and Xavier Giró Nieto. Budget-aware semi-supervised semantic and instance segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2019*, pages 93–102, 2019. 7

[6] Lyndon Chan, Mahdi S Hosseini, Corwyn Rowsell, Konstantinos N Plataniotis, and Savvas Damaskinos. Histosegnet: Semantic segmentation of histological tissue type in whole slide images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10662–10671, 2019. 2

[7] Bowen Cheng, Maxwell D Collins, Yukun Zhu, Ting Liu, Thomas S Huang, Hartwig Adam, and Liang-Chieh Chen. Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12475–12485, 2020. 2, 3, 6

[8] Junsuk Choe, Seungho Lee, and Hyunjung Shim. Attention-based dropout layer for weakly supervised single object localization and semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 8

[9] Hisham Cholakkal, Guolei Sun, Fahad Shahbaz Khan, and Ling Shao. Object counting and instance segmentation with image-level supervision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 12397–12405, 2019. 6

[10] Jifeng Dai, Kaiming He, and Jian Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1635–1643, 2015. 8

[11] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 2, 6

[12] Ruochen Fan, Ming-Ming Cheng, Qibin Hou, Tai-Jiang Mu, Jingdong Wang, and Shi-Min Hu. S4net: Single stage salient-instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6103–6112, 2019. 1, 3, 6

[13] 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 *Proceedings of the European conference on computer vision (ECCV)*, pages 367–383, 2018. 3, 6

[14] Weifeng Ge, Sheng Guo, Weilin Huang, and Matthew R Scott. Label-penet: Sequential label propagation and enhancement networks for weakly supervised instance segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3345–3354, 2019. 3, 6

[15] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 2, 6

[16] Lifeng He, Yuyan Chao, Kenji Suzuki, and Kesheng Wu. Fast connected-component labeling. *Pattern recognition*, 42(9):1977–1987, 2009. 4

[17] Seunghoon Hong, Donghun Yeo, Suha Kwak, Honglak Lee, and Bohyung Han. Weakly supervised semantic segmentation using web-crawled videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7322–7330, 2017. 8

[18] Cheng-Chun Hsu, Kuang-Jui Hsu, Chung-Chi Tsai, Yen-Yu Lin, and Yung-Yu Chuang. Weakly supervised instance segmentation using the bounding box tightness prior. *Advances in Neural Information Processing Systems*, 32:6586–6597, 2019. 8

[19] Jaedong Hwang, Seohyun Kim, Jeany Son, and Bohyung Han. Weakly supervised instance segmentation by deep community learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1020–1029, 2021. 3, 6

[20] Anna Khoreva, Rodrigo Benenson, Jan Hosang, Matthias Hein, and Bernt Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 876–885, 2017. 8

[21] Beomyoung Kim, Sangeun Han, and Junmo Kim. Discriminative region suppression for weakly-supervised semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1754–1761, 2021. 1, 2, 4

[22] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[23] Issam H Laradji, Negar Rostamzadeh, Pedro O Pinheiro, David Vazquez, and Mark Schmidt. Proposal-based instance segmentation with point supervision. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 2126–2130. IEEE, 2020. 6

[24] Issam H Laradji, David Vázquez, and Mark Schmidt. Where are the masks: Instance segmentation with image-level supervision. In *BMVC*, 2019. 6, 7

[25] Jungbeom Lee, Jooyoung Choi, Jisoo Mok, and Sungroh Yoon. Reducing information bottleneck for weakly super-

vised semantic segmentation. *Advances in Neural Information Processing Systems*, 34, 2021. 8

[26] Jungbeom Lee, Eunji Kim, Sungmin Lee, Jangho Lee, and Sungroh Yoon. Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5267–5276, 2019. 8

[27] Jungbeom Lee, Jihun Yi, Chaehun Shin, and Sungroh Yoon. Bbam: Bounding box attribution map for weakly supervised semantic and instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2643–2652, 2021. 8

[28] Seungho Lee, Minhyun Lee, Jongwuk Lee, and Hyunjung Shim. Railroad is not a train: Saliency as pseudo-pixel supervision for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5495–5505, 2021. 1

[29] Yi Li, Zhanghui Kuang, Liyang Liu, Yimin Chen, and Wayne Zhang. Pseudo-mask matters in weakly-supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6964–6973, 2021. 1, 2, 6, 7, 8

[30] Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3159–3167, 2016. 8

[31] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 2, 6

[32] Wei Liu, Andrew Rabinovich, and Alexander C Berg. Parsenet: Looking wider to see better. *arXiv preprint arXiv:1506.04579*, 2015. 6

[33] Yun Liu, Yu-Huan Wu, Pei-Song Wen, Yu-Jun 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*, 2020. 1, 3, 6, 7

[34] Davy Neven, Bert De Brabandere, Marc Proesmans, and Luc Van Gool. Instance segmentation by jointly optimizing spatial embeddings and clustering bandwidth. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8837–8845, 2019. 2

[35] Adrien Nivaggioli and Hicham Randrianarivo. Weakly supervised semantic segmentation of satellite images. In *2019 Joint urban remote sensing event (JURSE)*, pages 1–4. IEEE, 2019. 2

[36] Seong Joon Oh, Rodrigo Benenson, Anna Khoreva, Zeynep Akata, Mario Fritz, and Bernt Schiele. Exploiting saliency for object segmentation from image level labels. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pages 5038–5047. IEEE, 2017. 8

[37] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037, 2019. 7

[38] Jordi Pont-Tuset, Pablo Arbelaez, Jonathan T Barron, Ferran Marques, and Jitendra Malik. Multiscale combinatorial grouping for image segmentation and object proposal generation. *IEEE transactions on pattern analysis and machine intelligence*, 39(1):128–140, 2016. 1, 3, 6

[39] Yunhang Shen, Rongrong Ji, Yan Wang, Yongjian Wu, and Liujuan Cao. Cyclic guidance for weakly supervised joint detection and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 697–707, 2019. 6

[40] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5693–5703, 2019. 6

[41] Zhi Tian, Chunhua Shen, Xinlong Wang, and Hao Chen. Boxinst: High-performance instance segmentation with box annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5443–5452, 2021. 8

[42] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2):154–171, 2013. 3, 6

[43] Xiang Wang, Shaodi You, Xi Li, and Huimin Ma. Weakly-supervised semantic segmentation by iteratively mining common object features. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1354–1362, 2018. 8

[44] Yunchao Wei, Jiashi Feng, Xiaodan Liang, Ming-Ming Cheng, Yao Zhao, and Shuicheng Yan. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1568–1576, 2017. 2

[45] Bingfeng Zhang, Jimin Xiao, Yunchao Wei, Mingjie Sun, and Kaizhu Huang. Reliability does matter: An end-to-end weakly supervised semantic segmentation approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12765–12772, 2020. 1, 2, 7, 8

[46] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016. 1, 2, 4

[47] Yanzhao Zhou, Yi Zhu, Qixiang Ye, Qiang Qiu, and Jianbin Jiao. Weakly supervised instance segmentation using class peak response. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3791–3800, 2018. 1, 2, 4, 6

[48] Yi Zhu, Yanzhao Zhou, Huijuan Xu, Qixiang Ye, David Doermann, and Jianbin Jiao. Learning instance activation maps for weakly supervised instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3116–3125, 2019. 6