2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) | 978-1-6654-4509-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/CVPR46437.2021.00545

Railroad is not a Train: Saliency as Pseudo-pixel Supervision for Weakly Supervised Semantic Segmentation

Seungho Lee*
Yonsei University
seungholee@yonsei.ac.kr

Minhyun Lee*
Yonsei University
lmh315@yonsei.ac.kr

Jongwuk Lee
Sungkyunkwan University
jongwuklee@skku.edu

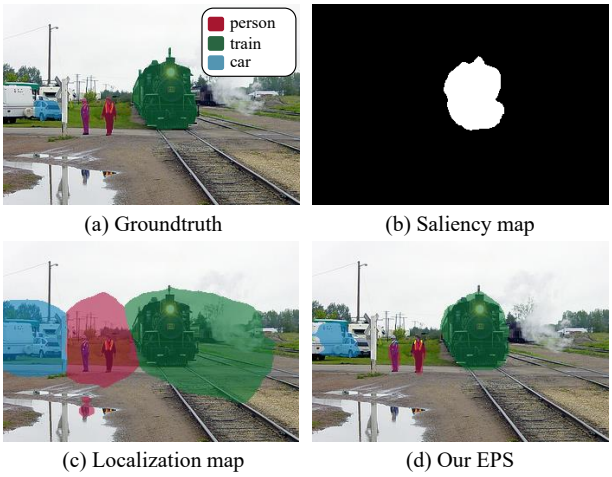
Hyunjung Shim†
Yonsei University
kateshim@yonsei.ac.kr

Abstract

Existing studies in weakly-supervised semantic segmentation (WSSS) using image-level weak supervision have several limitations: sparse object coverage, inaccurate object boundaries, and co-occurring pixels from non-target objects. To overcome these challenges, we propose a novel framework, namely Explicit Pseudo-pixel Supervision (EPS), which learns from pixel-level feedback by combining two weak supervisions; the image-level label provides the object identity via the localization map and the saliency map from the off-the-shelf saliency detection model offers rich boundaries. We devise a joint training strategy to fully utilize the complementary relationship between both information. Our method can obtain accurate object boundaries and discard co-occurring pixels, thereby significantly improving the quality of pseudo-masks. Experimental results show that the proposed method remarkably outperforms existing methods by resolving key challenges of WSSS and achieves the new state-of-the-art performance on both PASCAL VOC 2012 and MS COCO 2014 datasets. The code is available at https://github.com/halbielee/EPS.

1. Introduction

Weakly-supervised semantic segmentation (WSSS) utilizes weak supervision (*e.g.*, image-level labels [36, 37], scribbles [29], or bounding boxes [22]) and aims at achieving competitive performances to the fully-supervised model, which requires pixel-level labels. Most existing studies adopt image-level labels as the weak supervision of the segmentation model. The overall pipeline of WSSS consists of two stages. Firstly, pseudo-masks are generated for target objects using an image classifier. Then, the segmentation model is trained using the pseudo-masks as supervision. The prevalent technique for generating pseudo-masks is class activation mapping (CAM) [52], which provides object localization maps corresponding to their image-level labels. Because of the supervision gap between the fully (*i.e.*, pixel-level annotations) and weakly (*i.e.*, image-level labels) supervised semantic segmentation, WSSS has the following key challenges: 1) the localization map only captures a small fraction of target objects [52], 2) it suffers from the boundary mismatch of the objects [23], and 3) it hardly separates co-occurring pixels from target objects (*e.g.*, the railroad from a train) [25].



(a) Groundtruth (b) Saliency map

(c) Localization map (d) Our EPS

Figure 1. Motivating example of utilizing both the saliency map and the localization map for WSSS. (a) Groundtruth, (b) saliency map via PFAN [51], (c) localization map via CAM [52] and (d) our EPS utilizing both the saliency map and the localization map for training a classifier. Note that the saliency map cannot capture *person* and *car* while our result can correctly restore them, and the localization map overly captures two objects.

*indicates an equal contribution.
†Hyunjung Shim is a corresponding author.

978-1-6654-4509-2/21/$31.00 ©2021 IEEE
DOI 10.1109/CVPR46437.2021.00545

To address these problems, existing studies can be categorized into three pillars. The first approach expands object coverage to capture the full extent of objects by erasing pixels [9, 23, 28], ensembling score maps [21, 27], or using self-supervised signal [41]. However, they fail to determine accurate object boundaries of the target object because they have no clue to guide the object's shape. The second approach focuses on improving the object boundaries of pseudo-masks [13, 32]. Since they effectively learn object boundaries, they naturally expand pseudo-masks until boundaries. However, they still fail to distinguish coincident pixels of non-target objects from a target object. It is because the strong correlation between the foreground and the background (*i.e.*, co-occurrence) is almost indistinguishable from an inductive bias (*i.e.*, the frequency of observing the target object and its coincident pixels), as shown in [10]. Lastly, the third approach aims to mitigate the co-occurrence problem using extra groundtruth masks [24], or the saliency map [35, 47]. However, [24, 28] require strong pixel-level annotations, which are far from a weakly supervised learning paradigm. [35] is sensitive to the errors of the saliency map. Also, [47] does not cover the full extent of objects and suffers from the boundary mismatch.

In this paper, our goal is to overcome the three challenges of WSSS by fully utilizing both the localization map (*i.e.*, CAM from the image classifier trained with image-level labels) and the saliency map (*i.e.*, the output of the off-the-shelf saliency detection model [18, 34, 51]). We focus on a complementary relationship in the localization map and the saliency map. As illustrated in Figure 1, the localization map can distinguish different objects but does not separate their boundaries effectively. Contrarily, while the saliency map provides rich boundary information, it does not reveal object identity. In this sense, we argue that our method using two complementary pieces of information can resolve the performance bottleneck of WSSS.

To this end, we propose a novel framework for WSSS, called *Explicit Pseudo-pixel Supervision (EPS)*. To fully utilize the saliency map (*i.e.*, both the foreground and the background), we design a classifier to predict $C+1$ classes, consisting of C target classes and the background class. We leverage C localization maps and the background localization map to estimate a saliency map. Then, the saliency loss is defined as the pixel-wise difference between the saliency map and our estimated saliency map. By introducing the saliency loss, the model can be supervised by pseudo-pixel feedback across all classes. We also use the multi-label classification loss to predict image-level labels. Therefore, we train the classifier to optimize both the saliency loss and the multi-label classification loss, synergizing the predictions for both the background and foreground pixels– we find that our strategy can improve both the saliency map (Section 3.3 and Figure 3) and the pseudo-mask (Section 5.1 and Figure 4).

We stress that, because the saliency loss penalizes boundary mismatches via pseudo-pixel feedback, it can enforce our method to learn the object's accurate boundaries. As a byproduct, we can also capture the entire object by expanding the map until the boundaries. Because the saliency loss helps separate the foreground (*e.g.*, a train) from the background, our method can assign the co-occurring pixels (*e.g.*, a railroad) to the background class. Experimental results show that our EPS achieves remarkable segmentation performances, recording new state-of-the-art accuracies on PASCAL VOC 2012 and MS COCO 2014 datasets.

2. Related Work

Weakly-supervised semantic segmentation. The general pipeline of WSSS is to generate pseudo-masks from a classification network and to use the pseudo-masks as supervision to train a segmentation network. Due to the scarcity of boundary information in the image-level label, many existing methods suffer from inaccurate pseudo-masks. To address this problem, cross-image affinity [15], knowledge graph [31] and contrastive optimization [38, 50] are used to improve the quality of pseudo-masks. [5] proposes a self-supervised task to discover sub-categories to enforce the classifier to improve CAM. [1, 2] implicitly exploit the boundary information by calculating affinities between pixels. [49] focuses on producing reliable pixel-level annotations and designs an end-to-end network for generating segmentation maps. [20, 25] train the segmentation network by utilizing a boundary loss. Recently, [3] uses a single segmentation-based model with a self-supervised training scheme. [14] focuses on the robustness of the segmentation network by utilizing multiple incomplete pseudo-masks.

Saliency-guided semantic segmentation. Saliency detection (SD) methods generate the saliency map that distinguishes between the foreground and the background in an image via external saliency datasets with pixel-level annotations [18, 46, 51], or image-level annotations [39]. Many WSSS methods [15, 20, 27, 28, 42, 44] exploit the saliency map as the background cues of pseudo-masks. [43] utilizes the saliency map as the full supervision of single-object images. [16] uses an instance-level saliency map to learn the similarity graph for objects. [6, 40, 47] combine saliency maps with class-specific attention cues to generate reliable pseudo-masks. [48] jointly solves WSSS and SD using a single network to improve the performance of both tasks. Our EPS can be categorized into the saliency-guided method but is clearly distinguished from all others in the following reason. Most existing methods exploit the saliency map as a part of pseudo-masks or as the implicit guidance for refining the intermediate feature of the classifier. Contrarily, our method utilizes the saliency map as pseudo-pixel

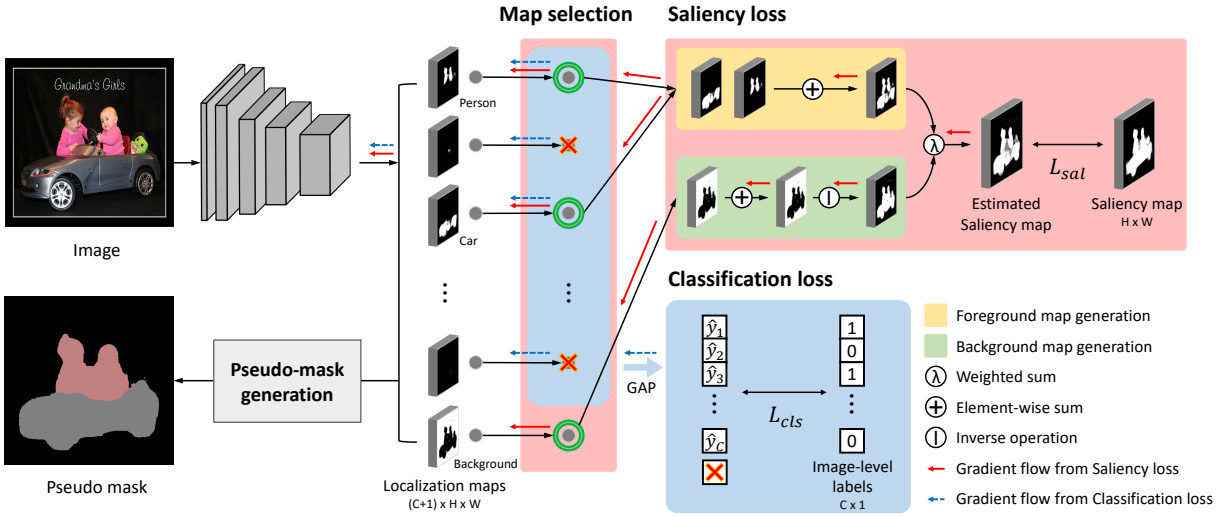


Figure 2. The overall framework of our EPS. $C+1$ localization maps are generated from a backbone network. The actual saliency map is generated from the off-the-shelf saliency detection model. Some localization maps for target labels are selectively used to generate an estimated saliency map (Section 3.2). The overall framework is jointly trained with the saliency loss and the classification loss (Section 3.3).

feedback for localization maps. Although [48] is the most similar work to ours in the sense of utilizing two complementary information, they neither address the co-occurring problem nor handle the noisy saliency map issue.

3. Proposed Method

In this section, we propose a new framework for Weakly-supervised semantic segmentation (WSSS), called *Explicit Pseudo-pixel Supervision (EPS)*. Considering two stages in WSSS, the first stage is to generate pseudo-masks and the second stage is to train the segmentation model. Here, our main contribution is to generate accurate pseudo-masks. Following the WSSS convention [13, 21, 27, 28, 41, 42], we then train a segmentation model, where the generated pseudo-masks in the first stage are used as supervision.

3.1. Motivation

Our key insight of EPS is to fully exploit two complementary information, *i.e.*, the object identity from the localization map and boundary information from the saliency map. To this end, we utilize the saliency map as pseudo-pixel feedback to the localization map for both target labels and the background. We devise a classifier with an additional background class, leading to predict a total of $C+1$ classes, as shown in Figure 2. Using the classifier, we can learn $C+1$ localization maps, *i.e.*, C localization maps for target labels and a background localization map.

We then explain how EPS can tackle both the boundary mismatch and co-occurrence problems in WSSS. To manage the boundary mismatch problem, we estimate the foreground map from C localization maps and match it with the foreground of the saliency map. In this way, the localization maps for target labels can receive pseudo-pixel feedback from the saliency map, thereby improving the boundaries of objects. To mitigate the co-occurring pixels of non-target objects, we also match the localization map for the background with the saliency map. Since the localization map for the background also receives pseudo-pixel feedback from the saliency map, the co-occurring pixels can be successfully assigned to the background; the co-occurring pixels of non-target objects mostly overlap with the background. It is why our method can separate the co-occurring pixels from target objects.

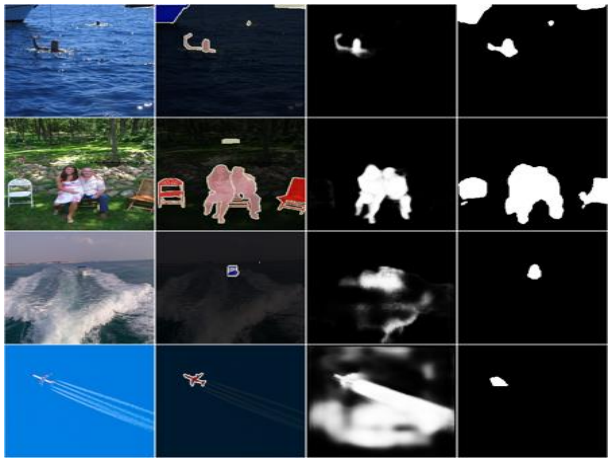


Figure 3. Qualitative examples of estimated saliency maps on PASCAL VOC 2012. (a) Input images, (b) groundtruth, (c) saliency maps from [51] and (d) our estimated saliency maps.

Lastly, the objective function of EPS is formulated with two parts: the saliency loss $\mathcal{L}_{sal}$ (marked by red box/arrow in Figure 2) via the saliency map, and the multi-label classification loss $\mathcal{L}_{cls}$ (marked by blue box/arrow in Figure 2)

via image-level labels. By jointly training the two objectives, we can synergize the localization map and the saliency map with complementary information– we observe that noisy and missing information of each other is complemented via our joint training strategy, as illustrated in Figure 3. For example, the original saliency map obtained from the off-the-shelf model [18, 34, 51] has missing and noisy information. On the other hand, our results successfully restore missing objects (*e.g.*, boats or chairs) and remove the noise (*e.g.*, water bubbles or contrail), which are evidently better than the original saliency map. Consequently, EPS can capture more accurate object boundaries and separate the co-occurring pixels from target objects. These advantages result in remarkable performance gains; Table 6 reports that EPS remarkably outperforms existing models up to 3.8–10.6% gains in terms of the segmentation accuracy.

3.2. Explicit Pseudo-pixel Supervision

We explain how to utilize the saliency map for pseudo-pixel supervision. The key advantage of the saliency map is to provide an object silhouette, which can better reveals object boundaries. To make use of this property, we match the saliency map with two cases: the foreground and the background. To make class-wise localization maps comparable with the saliency map, we merge the localization maps for target labels and generate a foreground map, $\mathbf{M}_{fg} \in \mathbb{R}^{H \times W}$. We can also represent the foreground by performing the inversion of a background map which is the localization map for the background label $\mathbf{M}_{bg} \in \mathbb{R}^{H \times W}$. (Later, we explain how to refine the foreground map to address noisy saliency maps.)

Specifically, we estimate the saliency map $\hat{\mathbf{M}}_s$ using $\mathbf{M}_{fg}$ and $\mathbf{M}_{bg}$ as follows:

$$\hat{\mathbf{M}}_s = \lambda \mathbf{M}_{fg} + (1 - \lambda)(1 - \mathbf{M}_{bg}), \tag{1}$$

where $\lambda \in [0, 1]$ is a hyperparameter to adjust a weighted sum of the foreground map and the inversion of the background map. (By default, we set λ to 0.5 in our experiments and an additional ablation study for λ is found in the supplementary material.) Then, we define the saliency loss $\mathcal{L}_{sal}$ as the sum of pixel-wise differences between our estimated saliency map and an actual saliency map. (The formal definition of $\mathcal{L}_{sal}$ is presented in Section 3.3.)

It is worth noting that using the pre-trained model is regarded as weakly supervised learning, thus utilizing the saliency map has been widely accepted as a common practice in WSSS. Despite its popularity, adopting the fully supervised saliency detection model can be arguable in that they use pixel-level annotations from different datasets. In this paper, we investigate the effect of different saliency detection methods; 1) unsupervised and 2) fully supervised saliency detection models (see Section 5.3), and empirically show our method using any of them outperforms all other methods [13, 21, 40, 43, 47] using fully supervised saliency models. Whereas existing methods are limited to fully take advantage of the saliency map, our method incorporates the saliency map as pseudo-pixel supervision and exploits it as the cues for boundaries and co-occurring pixels.

Map selection for handling saliency bias. Previously, we assume that the foreground map can be the union of the localization maps for target labels; the background map can be the localization map of the background label. However, such a naïve selection rule may not be compatible with the saliency map computed by the off-the-shelf model. For example, the saliency map from [51] often ignores some objects as salient objects (*e.g.*, small people nearby a train in Figure 1). This systematic error is inevitable because the saliency model learns the statistics of different datasets. Unless considering this error, the same error may propagate to our model and lead the performance degradation.

To tackle the systematic error, we develop an effective strategy using the overlapping ratio between the localization map and the saliency map. Specifically, the i-th localization map $\mathbf{M}_i$ is assigned to the foreground if $\mathbf{M}_i$ is overlapped with the saliency map more than τ%, otherwise the background. Formally, the foreground and the background map are computed by:

$$\begin{aligned} \mathbf{M}_{fg} &= \sum_{i=1}^{C} y_i \cdot \mathbf{M}_i \cdot \mathbb{1}[\mathcal{O}(\mathbf{M}_i, \mathbf{M}_s) > \tau], \\ \mathbf{M}_{bg} &= \sum_{i=1}^{C} y_i \cdot \mathbf{M}_i \cdot \mathbb{1}[\mathcal{O}(\mathbf{M}_i, \mathbf{M}_s) \leq \tau] + \mathbf{M}_{C+1}, \end{aligned} \tag{2}$$

where $y \in \mathbb{R}^C$ is the binary image-level label and $\mathcal{O}(\mathbf{M}_i, \mathbf{M}_s)$ is the function to compute the overlapping ratio between $\mathbf{M}_i$ and $\mathbf{M}_s$. For that, we first binarize the localization map and the saliency map such that: for pixel p, $\mathbf{B}_k(p) = 1$ if $\mathbf{M}_k(p) > 0.5$; $\mathbf{B}_k(p) = 0$, otherwise. $\mathbf{B}_i$ and $\mathbf{B}_s$ are the binarized maps corresponding to $\mathbf{M}_i$ and $\mathbf{M}_s$, respectively. We then compute the overlapping ratio between $\mathbf{M}_i$ and $\mathbf{M}_s$, *i.e.*, $\mathcal{O}(\mathbf{M}_i, \mathbf{M}_s) = |\mathbf{B}_i \cap \mathbf{B}_s| / |\mathbf{B}_i|$. We set $\tau = 0.4$ regardless of datasets and backbone models. In the supplementary material, we show that our method is robust against the choice of τ (*i.e.*, τ within [0.3, 0.5] shows the comparable performance).

Instead of a single localization map for the background label, we combine the localization map for the background label with the localization maps not selected as the foreground. Although it is simple, we can bypass the error of the saliency map and effectively train some objects neglected from the saliency map. (In Table 3, we report the effectiveness of the proposed strategy to overcome the error of the saliency map.)

3.3. Joint Training Procedure

Using the saliency map and image-level labels, the overall training objective of EPS consists of two parts, the

saliency loss $\mathcal{L}_{sal}$ and the classification loss $\mathcal{L}_{cls}$. First, the saliency loss $\mathcal{L}_{sal}$ is formulated by measuring the average pixel-level distance between the actual saliency map $\mathbf{M}_s$ and the estimated saliency map $\hat{\mathbf{M}}_s$.

$$\mathcal{L}_{sal} = \frac{1}{H \cdot W} ||\mathbf{M}_s - \hat{\mathbf{M}}_s||^2, \quad (3)$$

where $\mathbf{M}_s$ is obtained from the off-the-shelf saliency detection model– PFAN [51] trained on DUTS dataset [39]. Note that our method consistently outperforms all previous arts regardless of the saliency detection models.

Next, the classification loss is computed by a multi-label soft margin loss between the image-level label y and its prediction $\hat{y} \in \mathbb{R}^C$, which is the result of the global average pooling on the localization map for each target class.

$$\mathcal{L}_{cls} = -\frac{1}{C} \sum_{i=1}^{C} y_i \log \sigma(\hat{y}_i) + (1 - y_i) \log (1 - \sigma(\hat{y}_i)), \quad (4)$$

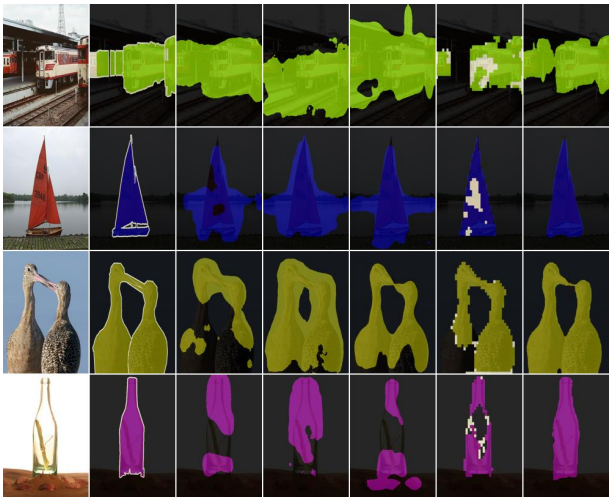
where $\sigma(\cdot)$ is the sigmoid function. Finally, the total training loss is the sum of the multi-label classification loss and the saliency loss, *i.e.*, $\mathcal{L}_{total} = \mathcal{L}_{cls} + \mathcal{L}_{sal}$.

As shown in Figure 2, $\mathcal{L}_{sal}$ is involved in updating the parameters of $C + 1$ classes, including target objects and the background. Meanwhile, $\mathcal{L}_{cls}$ only evaluates the label prediction for C classes, excluding the background class– the gradient from $\mathcal{L}_{cls}$ does not flow into the background class. However, the prediction of the background class can be implicitly affected by $\mathcal{L}_{cls}$ because it supervises classifier training.

4. Experimental Setup

Datasets. We conduct an empirical study on two popular benchmark datasets, PASCAL VOC 2012 [12] and MS COCO 2014 [30]. PASCAL VOC 2012 consists of 21 classes (*i.e.*, 20 objects and the background) with 1,464, 1,449, and 1,456 images for training, validation, and test set, respectively. Following the common practice in semantic segmentation, we use the augmented training set with 10,582 images [17]. Next, COCO 2014 consists of 81 classes, including a background, with 82,081 and 40,137 images for training and validation, where images with no target classes are excluded as done in [9]. Because the groundtruth segmentation labels of some objects overlap each other, we adopt the groundtruth segmentation labels from COCO-Stuff [4], which solves the overlapping problem on the same COCO dataset.

Evaluation protocol. We validate our method with the validation and the test set on PASCAL VOC 2012, and the validation set on COCO 2014. The evaluation results on the test set of PASCAL VOC 2012 is obtained from the official PASCAL VOC evaluation server. Also, we adopt mean intersection-over-union (mIoU) to measure the accuracy of segmentation models.



(a) (b) (c) (d) (e) (f) (g)

Figure 4. Qualitative comparison for pseudo-masks on PASCAL VOC 2012. (a) Input images, (b) groundtruth, (c) CAM, (d) SEAM, (e) ICD, (f) SGAN and (g) our EPS.

Implementation details. We choose ResNet38 [45] as the backbone network of our method with the output stride of 8. All backbone models are pre-trained on ImageNet [11]. We use the SGD optimizer with a batch size of 8. Our method is trained until 20k iterations with learning rate 0.01 (0.1 for the last convolutional layer). For data augmentation, we use a random scaling, random flipping, and random crop into 448×448. For the segmentation networks, we adopt DeepLab-LargeFOV (V1) [7] and DeepLab-ASPP (V2) [8], and VGG16 and ResNet101 for their backbone networks. Specifically, we use four segmentation networks: VGG16-based DeepLab-V1 and DeepLab-V2, ResNet101 based DeepLab-V1 and DeepLab-V2. More detailed setting is in the supplementary material.

5. Experimental Results

5.1. Handling Boundary and Co-occurrence

Boundary mismatch problem. To validate the boundary of pseudo-masks, we compare the quality of boundaries with the state-of-the-art methods [32, 41, 52]. We utilize SBD [17], which provides boundary annotations and the boundary benchmark in PASCAL VOC 2011. As done in [32], the quality of the boundary is evaluated in a class-agnostic manner by computing the edges of pseudo-masks from the Laplacian edge detector. Then, the boundary quality is evaluated by measuring recall, precision, and F1-score, comparing the predicted and groundtruth boundaries. Table 1 reports that our method largely outperforms other methods in all three metrics. The qualitative examples in Figure 4 show that our method can capture more accurate boundaries than all the other methods.

Co-occurrence problem. As discussed in several stud-

Method	Recall (%)	Precision (%)	F1-score (%)
CAM [52]$_{\text{CVPR'16}}$	22.3	35.8	27.5
SEAM [41]$_{\text{CVPR'20}}$	40.2	45.0	42.5
BES [32]$_{\text{ECCV'20}}$	45.5	46.4	45.9
Our EPS	60.0	73.1	65.9

Table 1. Boundary accuracy evaluated on the SBD trainval set. Note that the results of BES are measured from the boundary prediction network proposed in [32].

Method	*boat* w/ *water*	*train* w/ *railroad*	*train* w/ *platform*
CAM [52]$_{\text{CVPR'16}}$	0.74 (33.1)	0.11 (52.9)	0.09 (49.6)
SEAM [41]$_{\text{CVPR'20}}$	1.13 (30.7)	0.24 (48.6)	0.20 (45.5)
ICD [13]$_{\text{CVPR'20}}$	0.47 (41.4)	0.11 (56.7)	0.09 (49.2)
SGAN [47]$_{\text{ACCESS'20}}$	0.10 (42.3)	0.02 (48.8)	0.01 (36.3)
Our EPS	0.10 (55.0)	0.02 (78.1)	0.01 (73.0)

Table 2. Comparison with representative existing methods handling the co-occurrence problem. Each entry is $m_{k,c}$ in blue (the lower the better) and IoU in the bracket (the higher the better).

ies [20, 25, 28, 35], we observe that some background classes frequently appear with target objects in PASCAL VOC 2012. We quantitatively analyze the frequency of co-occurred objects by employing the PASCAL-CONTEXT dataset [33], which provides pixel-level annotations for a whole scene (*e.g.*, *water* and *railroad*). We choose three co-occurring pairs; *boat* with *water*, *train* with *railroad*, and *train* with *platform*. We compare IoU for the target class and the *confusion ratio* between a target class and its coincident class. The confusion ratio measures how much the coincident class is incorrectly predicted as the target class. The confusion ratio $m_{k,c}$ is calculated by $m_{k,c} = FP_{k,c}/TP_c$, where $FP_{k,c}$ is the number of pixels mis-classified as the target class c for the coincident class k, and TP_c is the number of true-positive pixels for the target class c. More detailed analysis on the co-occurrence problem is in the supplementary materials.

Table 2 reports that EPS consistently shows a lower confusion ratio than other methods. SGAN [47] has quite a similar confusion ratio with ours, but our method captures the target class much accurately in terms of IoU. Interestingly, SEAM shows a high confusion ratio and even worse than CAM. It is because SEAM [41] learns to cover the full extent of target objects by applying self-supervised training, which is easily fooled by the coincident pixels of target objects. Meanwhile, CAM only captures the most discriminative region of target objects and does not cover the less discriminative parts, *e.g.*, the coincident class. We can also observe this phenomenon in Figure 4.

	Baseline	Naïve	Pre-defined	Our adaptive
mIoU	66.1	66.5	67.9	69.4

Table 3. Effect of map selection strategies. The accuracies of pseudo-masks using different map selection strategies are evaluated on the PASCAL VOC 2012 train set.

Method	w/o refinement	w/ CRF [26]	w/ AffinityNet [2]
CAM [52]$_{\text{CVPR'16}}$	48.0	-	58.1
SEAM [41]$_{\text{CVPR'20}}$	55.4	56.8	63.6
ICD [32]$_{\text{CVPR'20}}$*	59.9	62.2	-
SGAN [47]$_{\text{ACCESS'20}}$*	62.8	-	-
Our EPS	69.4	71.4	71.6

Table 4. Accuracy (mIoU) of pseudo-masks evaluated on the PASCAL VOC 2012 train set. Note that * indicates that low-confident pixels are ignored; other methods use all pixels for evaluation.

5.2. Effect of Map Selection Strategies

We evaluate the effectiveness of our map selection strategy to mitigate the error of the saliency map. We compare three different map selection strategies to the baseline, which does not use the map selection module. As the naïve strategy, the foreground map is the union of all object localization maps; the background map equals the localization map of the background class (*i.e.*, naïve strategy). Next, we follow the naïve strategy with the following exceptions. The localization maps of several pre-determined classes (*e.g.*, *sofa*, *chair*, and *dining table*) are assigned to the background map (*i.e.*, pre-defined class strategy). Lastly, the proposed selection method utilizes the overlapping ratio between the localization map and the saliency map, as explained in Section 3.2 (*i.e.*, our adaptive strategy).

Table 3 shows that our adaptive strategy can effectively handle the systematic bias of the saliency map. The naïve strategy implies no bias consideration when generating the estimated saliency map from the localization maps. In this case, the performance of pseudo-masks is degraded, especially on *sofa*, *chair* or *dining table* classes. The performance of using pre-defined classes shows that the bias can be mitigated by neglecting missing classes in the saliency map. However, as it requires manual selection by human observers, it is less practical and cannot make an optimal decision per image. Meanwhile, our adaptive strategy can handle the bias automatically and makes more effective decisions for a given saliency map.

5.3. Comparison with state-of-the-arts

Accuracy of pseudo-masks. We adopt a multi-scale inference by aggregating the prediction results from images with different scales, which is a common practice utilized in [2, 41]. Then, We evaluate the accuracies of pseudo-masks in the train set by comparing our EPS with the

Figure 5. Qualitative examples of segmentation results on PASCAL VOC 2012. (a) Input images, (b) groundtruth and (c) our EPS.

baseline CAM [52] and three state-of-the-art methods, *i.e.*, SEAM [41], ICD [13], and SGAN [47]. Here, measuring the accuracy of the pseudo-masks in the train set is a common protocol in WSSS because the pseudo-masks of the train set are used to supervise the segmentation model. Table 4 summarizes the accuracies of pseudo-masks and indicates that our method clearly outperforms all existing methods by large margins (*i.e.*, 7–21% gaps). Figure 4 visualizes the qualitative examples of pseudo-masks, confirming that our method remarkably improves the object boundary and significantly outperforms three state-of-the-art methods in terms of the quality of pseudo-masks. Our method can capture the precise boundaries of objects (2nd row) and thus naturally cover the full extent of objects (3rd row), and also mitigate the coincident pixels (1st row). More examples and failure cases of our method are provided in the supplementary material.

Accuracy of segmentation maps. Previous methods [2, 13, 41] generate pseudo-masks and refine them with the CRF post-processing algorithm [26] or affinity network [2]. Meanwhile, as shown in Table 4, our generated pseudo-masks are accurate enough, thereby we train a segmentation network without any additional refinement for pseudo-masks. We extensively evaluate and precisely compare our method with others on the four segmentation networks in the Pascal VOC 2012 dataset.

Our method performs remarkably better than other methods regardless of segmentation networks. Table 5 reports that our method is more accurate than other methods with the same VGG16 backbone. Besides, our results on the VGG16 are comparable or even superior to other existing methods based on a more powerful backbone (*i.e.* ResNet101 in Table 6). Our method also shows a clear improvement over existing methods. Finally, Table 6 demonstrates that our method (under ResNet101 based DeepLab-V1 with saliency map) achieves the new state-of-the-art performance (71.0 for validation and 71.8 for test set) in the PASCAL VOC 2012 dataset. We highlight that the gains

Method	Seg.	Sup.	val	test
SEC [25]$_{\text{ECCV'16}}$	V1	I.	50.7	51.7
AffinityNet [2]$_{\text{CVPR'18}}$	V1	I.	58.4	60.5
ICD [13]$_{\text{CVPR'20}}$	V1	I.	61.2	60.9
BES [32]$_{\text{ECCV'20}}$	V1	I.	60.1	61.1
GAIN [28]$_{\text{CVPR'18}}$	V1	I.+S.	55.3	56.8
MCOF [40]$_{\text{CVPR'18}}$	V1	I.+S.	56.2	57.6
SSNet [48]$_{\text{ICCV'19}}$	V1	I.+S.	57.1	58.6
DSRG [20]$_{\text{CVPR'18}}$	V2	I.+S.	59.0	60.4
SeeNet [19]$_{\text{NeurIPS'18}}$	V1	I.+S.	61.1	60.7
MDC [44]$_{\text{CVPR'18}}$	V1	I.+S.	60.4	60.8
FickleNet [27]$_{\text{CVPR'18}}$	V2	I.+S.	61.2	61.9
OAA [21]$_{\text{ICCV'19}}$	V1	I.+S.	63.1	62.8
ICD [13]$_{\text{CVPR'20}}$	V1	I.+S.	64.0	63.9
Multi-Est. [14]$_{\text{ECCV'20}}$	V1	I.+S.	64.6	64.2
Split. & Merge. [50]$_{\text{ECCV'20}}$	V2	I.+S.	63.7	64.5
SGAN [47]$_{\text{ACCESS'20}}$	V2	I.+S.	64.2	65.0
Our EPS	V1	I.+S.	66.6	**67.9**
	V2	I.+S.	**67.0**	67.3

Table 5. Segmentation results (mIoU) on PASCAL VOC 2012. All results are based on VGG16. The best score is in bold throughout all experiments.

achieved by the existing state-of-the-art models were approximately 1%. Meanwhile, our method achieves more than 3% higher gains than the previous best record. Figure 5 visualizes the qualitative examples of our segmentation results on PASCAL VOC 2012. These results confirm that our method provides accurate boundaries and successfully resolves the co-occurrence problem.

In Table 7, we further evaluate our method in the COCO 2014 dataset. We use VGG16 based DeepLab-V2 as the segmentation network to compare with SGAN [47], which is the state-of-the-art WSSS model in the COCO dataset. Our method achieves 35.7 mIoU in the validation set, and it is 1.9% higher than SGAN [47]. Consequently, we achieve the new state-of-the-art accuracy in the COCO 2014 dataset. These outstanding performances over the existing state-of-the-arts on both datasets confirm the effectiveness of our

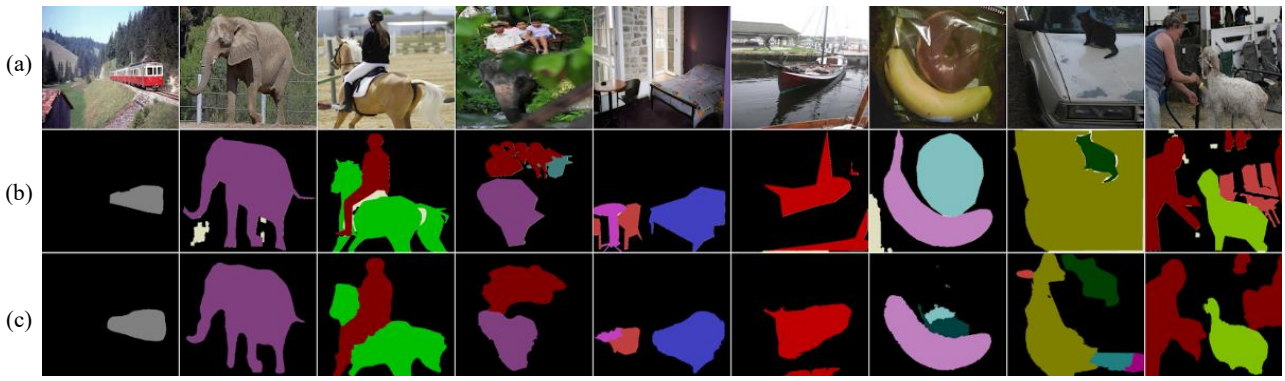


Figure 6. Qualitative examples of segmentation results on MS COCO 2014. (a) Input images, (b) groundtruth and (c) our EPS.

Method	Seg.	Sup.	val	test
ICD [13]$_{\text{CVPR'20}}$	V1	I.	64.1	64.3
SC-CAM [5]$_{\text{CVPR'20}}$	V1	I.	66.1	65.9
BES [32]$_{\text{ECCV'20}}$	V2	I.	65.7	66.6
LIID [31]$_{\text{TPAMI'20}}$	V2	I.	66.5	67.5
MCOF [40]$_{\text{CVPR'18}}$	V1	I.+S.	60.3	61.2
SeeNet [19]$_{\text{NeurIPS'18}}$	V1	I.+S.	63.1	62.8
DSRG [20]$_{\text{CVPR'18}}$	V2	I.+S.	61.4	63.2
FickleNet [27]$_{\text{CVPR'18}}$	V2	I.+S.	64.9	65.3
OAA [21]$_{\text{ICCV'19}}$	V1	I.+S.	65.2	66.4
Multi-Est. [14]$_{\text{ECCV'19}}$	V1	I.+S.	67.2	66.7
MCIS [38]$_{\text{ECCV'20}}$	V1	I.+S.	66.2	66.9
SGAN [47]$_{\text{ACCESS'20}}$	V2	I.+S.	67.1	67.2
ICD [13]$_{\text{CVPR'20}}$	V1	I.+S.	67.8	68.0
Our EPS	V1	I.+S.	**71.0**	**71.8**
	V2	I.+S.	70.9	70.8

Table 6. Segmentation results (mIoU) on PASCAL VOC 2012. All results are based on ResNet101.

Method	Seg.	Sup.	val
SEC [25]$_{\text{ECCV'16}}$	V1	I.	22.4
DSRG [20]$_{\text{CVPR'18}}$	V2	I.+S.	26.0
ADL [9]$_{\text{TPAMI'20}}$	V1	I.+S.	30.8
SGAN [47]$_{\text{ACESS'20}}$	V2	I.+S.	33.6
Our EPS	V2	I.+S.	**35.7**

Table 7. Segmentation results (mIoU) on MS COCO 2014. All results are based on VGG16.

method; by fully utilizing both localization maps and the saliency map, it successfully captures the integral of target objects correctly and remedies the shortcomings of existing models. Figure 6 shows the qualitative examples of segmentation results on the COCO 2014 dataset. Our method performs well when a few objects appear without occlusions but less effective in handling many small objects. More examples and failure cases of our method are provided in the supplementary material.

Effect of saliency detection models. To investigate the effect of different saliency detection models, we adopt three saliency models; PFAN [51] (our default), DSS [18] used by OAA [21] and ICD [13], and USPS [34] (*i.e.*, the unsupervised detection model). The segmentation results (mIoU) under Resnet101 based DeepLab-V1 are 71.0/71.8 with PFAN, 70.0/70.1 with DSS, and 68.8/69.9 with USPS (validation set and test set), respectively. These scores support that our EPS using any of three different saliency models is still more accurate than all the other methods in Table 6. Notably, our EPS using the unsupervised saliency model outperforms all existing methods using the supervised saliency model.

6. Conclusion

We propose a novel weakly supervised segmentation framework, namely *explicit pseudo-pixel supervision (EPS)*. Motivated by the complementary relationship between the localization map and the saliency map, our EPS learns from pseudo-pixel feedback combining with the saliency map and the localization map. Owing to our joint training scheme, we successfully complement noise or missing information on both sides. Consequently, our EPS can capture precise object boundaries and discard co-occurring pixels of non-target objects, remarkably improving the quality of pseudo-masks. Extensive evaluations and various case studies demonstrate the effectiveness of our EPS and the outstanding performances, the new state-of-the-art accuracies for WSSS on both PASCAL VOC 2012 and MS COCO 2014 datasets.

Acknowledgements. We thank Duhyeon Bang and Junsuk Choe for the feedback. This research was supported by the Basic Science Research Program through the NRF Korea funded by the MSIP (NRF-2019R1A2C2006123, 2020R1A4A1016619), the IITP grant funded by the MSIT (2020-0-01361, Artificial Intelligence Graduate School Program (YONSEI UNIVERSITY)), and the Korea Medical Device Development Fund grant funded by the Korean government (Project Number: 202011D06).

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. 2

[2] Jiwoon Ahn and Suha Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4981–4990, 2018. 2, 6, 7

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

[4] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1209–1218, 2018. 5

[5] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8991–9000, 2020. 2, 8

[6] Arslan Chaudhry, Puneet K. Dokania, and Philip H. S. Torr. Discovering class-specific pixels for weakly-supervised semantic segmentation. In *Proceedings of the British Machine Vision Conference*, 2017. 2

[7] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. In *International Conference on Learning Representations*, 2015. 5

[8] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2017. 5

[9] 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. 2, 5, 8

[10] Junsuk Choe, Seong Joon Oh, Seungho Lee, Sanghyuk Chun, Zeynep Akata, and Hyunjung Shim. Evaluating weakly supervised object localization methods right. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3133–3142, 2020. 2

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, 2009. 5

[12] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, 2015. 5

[13] Junsong Fan, Zhaoxiang Zhang, Chunfeng Song, and Tieniu Tan. Learning integral objects with intra-class discriminator for weakly-supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4283–4292, 2020. 2, 3, 4, 6, 7, 8

[14] Junsong Fan, Zhaoxiang Zhang, and Tieniu Tan. Employing multi-estimations for weakly-supervised semantic segmentation. In *Proceedings of the European Conference on Computer Vision*, 2020. 2, 7, 8

[15] Junsong Fan, Zhaoxiang Zhang, Tieniu Tan, Chunfeng Song, and Jun Xiao. Cian: Cross-image affinity net for weakly supervised semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10762–10769, 2020. 2

[16] 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*, pages 367–383, 2018. 2

[17] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *2011 International Conference on Computer Vision*, pages 991–998. IEEE, 2011. 5

[18] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip HS Torr. Deeply supervised salient object detection with short connections. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3203–3212, 2017. 2, 4, 8

[19] Qibin Hou, PengTao Jiang, Yunchao Wei, and Ming-Ming Cheng. Self-erasing network for integral object attention. In *Advances in Neural Information Processing Systems*, pages 549–559, 2018. 7, 8

[20] Zilong Huang, Xinggang Wang, Jiasi Wang, Wenyu Liu, and Jingdong Wang. Weakly-supervised semantic segmentation network with deep seeded region growing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7014–7023, 2018. 2, 6, 7, 8

[21] Peng-Tao Jiang, Qibin Hou, Yang Cao, Ming-Ming Cheng, Yunchao Wei, and Hong-Kai Xiong. Integral object mining via online attention accumulation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2070–2079, 2019. 2, 3, 4, 7, 8

[22] 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. 1

[23] Dahun Kim, Donghyeon Cho, Donggeun Yoo, and In So Kweon. Two-phase learning for weakly supervised object localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3534–3543, 2017. 1, 2

[24] Alexander Kolesnikov and Christoph Lampert. Improving weakly-supervised object localization by micro-annotation. In Edwin R. Hancock Richard C. Wilson and William A. P. Smith, editors, *Proceedings of the British Machine Vision Conference*, pages 92.1–92.12. BMVA Press, September 2016. 2

[25] Alexander Kolesnikov and Christoph H Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *Proceedings of the European Conference on Computer Vision*, pages 695–711. Springer, 2016. 1, 2, 6, 7, 8

[26] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *Advances in Neural Information Processing Systems*, pages 109–117, 2011. 6, 7

[27] 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 Conference on Computer Vision and Pattern Recognition*, pages 5267–5276, 2019. 2, 3, 7, 8

[28] Kunpeng Li, Ziyan Wu, Kuan-Chuan Peng, Jan Ernst, and Yun Fu. Tell me where to look: Guided attention inference network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9215–9223, 2018. 2, 3, 6, 7

[29] 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. 1

[30] 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 *Proceedings of the European Conference on Computer Vision*, pages 740–755. Springer, 2014. 5

[31] 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. 2, 8

[32] Chen Liyi, Wu Weiwei, Chenchen Fu, Xiao Han, and Yuntao Zhang. Weakly supervised semantic segmentation with boundary exploration. In *Proceedings of the European Conference on Computer Vision*, 2020. 2, 5, 6, 7, 8

[33] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 891–898, 2014. 6

[34] Tam Nguyen, Maximilian Dax, Chaithanya Kumar Mummadi, Nhung Ngo, Thi Hoai Phuong Nguyen, Zhongyu Lou, and Thomas Brox. Deepusps: Deep robust unsupervised saliency prediction via self-supervision. In *Advances in Neural Information Processing Systems*, pages 204–214, 2019. 2, 4, 8

[35] 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*, pages 5038–5047. IEEE, 2017. 2, 6

[36] Deepak Pathak, Philipp Krahenbuhl, and Trevor Darrell. Constrained convolutional neural networks for weakly supervised segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1796–1804, 2015. 1

[37] Pedro O Pinheiro and Ronan Collobert. From image-level to pixel-level labeling with convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1713–1721, 2015. 1

[38] Guolei Sun, Wenguan Wang, Jifeng Dai, and Luc Van Gool. Mining cross-image semantics for weakly supervised semantic segmentation. In *Proceedings of the European Conference on Computer Vision*, 2020. 2, 8

[39] Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to detect salient objects with image-level supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 136–145, 2017. 2, 5

[40] 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. 2, 4, 7, 8

[41] Yude Wang, Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12275–12284, 2020. 2, 3, 5, 6, 7

[42] 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, 3

[43] Yunchao Wei, Xiaodan Liang, Yunpeng Chen, Xiaohui Shen, Ming-Ming Cheng, Jiashi Feng, Yao Zhao, and Shuicheng Yan. Stc: A simple to complex framework for weakly-supervised semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(11):2314–2320, 2016. 2, 4

[44] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7268–7277, 2018. 2, 7

[45] Zifeng Wu, Chunhua Shen, and Anton Van Den Hengel. Wider or deeper: Revisiting the resnet model for visual recognition. *Pattern Recognition*, 90:119–133, 2019. 5

[46] Huaxin Xiao, Jiashi Feng, Yunchao Wei, Maojun Zhang, and Shuicheng Yan. Deep salient object detection with dense connections and distraction diagnosis. *IEEE Transactions on Multimedia*, 20(12):3239–3251, 2018. 2

[47] Qi Yao and Xiaojin Gong. Saliency guided self-attention network for weakly and semi-supervised semantic segmentation. *IEEE Access*, 8:14413–14423, 2020. 2, 4, 6, 7, 8

[48] Yu Zeng, Yunzhi Zhuge, Huchuan Lu, and Lihe Zhang. Joint learning of saliency detection and weakly supervised semantic segmentation. In *Proceedings of the IEEE International*

Conference on Computer Vision, pages 7223–7233, 2019. 2, 3, 7

[49] 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*, pages 12765–12772. AAAI Press, 2020. 2

[50] Tianyi Zhang, Guosheng Lin, Weide Liu, Jianfei Cai, and Alex Kot. Splitting vs. merging: Mining object regions with discrepancy and intersection loss for weakly supervised semantic segmentation. In *Proceedings of the European Conference on Computer Vision*, 2020. 2, 7

[51] Ting Zhao and Xiangqian Wu. Pyramid feature attention network for saliency detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3085–3094, 2019. 1, 2, 3, 4, 5, 8

[52] 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, 5, 6, 7