Small Objects Matters in Weakly-supervised Semantic Segmentation

Cheolhyun Mun[*][†]
Samsung Research
Seoul, Korea
cheolhyunmun@yonsei.ac.kr

Sanghuk Lee[*][†]
SOCAR AI Research
Seoul, Korea
li-xh16@yonsei.ac.kr

Youngjung Uh
Yonsei University
Seoul, Korea
yj.uh@yonsei.ac.kr

Junsuk Choe
Sogang University
Seoul, Korea
jschoe@sogang.ac.kr

Hyeran Byun
Yonsei University
Seoul, Korea
hrbyun@yonsei.ac.kr

Abstract

Weakly-supervised semantic segmentation (WSSS) performs pixel-wise classification given only image-level labels for training. Despite the difficulty of this task, the research community has achieved promising results over the last five years. Still, current WSSS literature misses the detailed sense of how well the methods perform on different sizes of objects. Thus we propose a novel evaluation metric to provide a comprehensive assessment across different object sizes and collect a size-balanced evaluation set to complement PASCAL VOC. With these two gadgets, we reveal that the existing WSSS methods struggle in capturing small objects. Furthermore, we propose a size-balanced cross-entropy loss coupled with a proper training strategy. It generally improves existing WSSS methods as validated upon ten baselines on three different datasets.

1. Introduction

Recently, weakly-supervised learning (WSL) has been attracting attention because of its low-cost annotation. Among many tasks, weakly-supervised semantic segmentation (WSSS) methods learn to predict semantic segmentation masks given only weak labels such as image-level class labels for training.

To solve this problem, existing WSSS techniques generate pseudo segmentation masks from a classification network and then train a fully-supervised semantic segmentation model such as DeepLabV2 [4]. To improve WSSS performances, most existing methods have focused on producing more accurate pseudo labels. With this strategy, WSSS

performances have been greatly improved in the last five years [1, 26, 29, 30, 35, 38, 42, 43, 45, 50].

However, we lack a detailed sense of performance: do methods with high mIoU always better capture all the details? Interestingly, we observe that some methods with lower mIoU better capture small objects than others. Although it is undoubtedly important that the segmentation model also correctly captures small objects, this limitation has not been well studied yet in WSSS literature. How does each method behave in different types of environments? To answer this question, we address the limitations of the conventional metric, the dataset, and the training objective, and propose a complement thereby we anticipate WSSS techniques to become more complete and applicable to different needs.

Conventional metric (mIoU) and its pitfall. mIoU is mean of per-class IoUs where IoU is the intersection-over-union of the segmented objects. While an IoU is depicted with one predicted segment and one ground-truth segment, it pre-accumulates *all* predicted pixels and *all* ground-truth pixels in the entire dataset (Fig. 2 (a)). mIoU has widely been used to measure the performance of different models in semantic segmentation.

Despite of its usefulness in measuring the overall accuracy of segmentation predictions, mIoU does not account for the comprehensiveness of the predictions. As illustrated in Fig. 1 (a), *Prediction 1* and *Prediction 2* have the same IoU score since they miss the same number of pixels. However, in *Prediction 1*, the red cross marks indicate a complete failure in object segmentation, while *Prediction 2* can be considered as minor errors.

Conventional dataset. The PASCAL VOC 2012 [13] is the representative benchmark for WSSS. The problem is, however, the evaluation set of VOC has an imbalanced distribution in terms of object-size. Fig. 1 (b) shows the over-

---

* indicates equal contribution
†The work was done while the author was at Yonsei University

Ground Truth **Prediction 1** **Prediction 2**

IoU 100%   Prediction 1 & 2 have same IoU (90%)

(a) Large object domination problem in mIoU

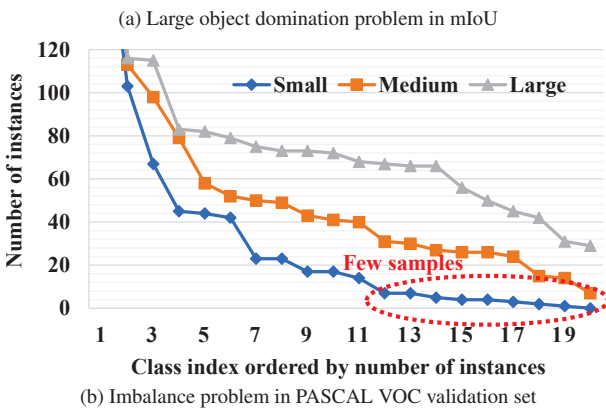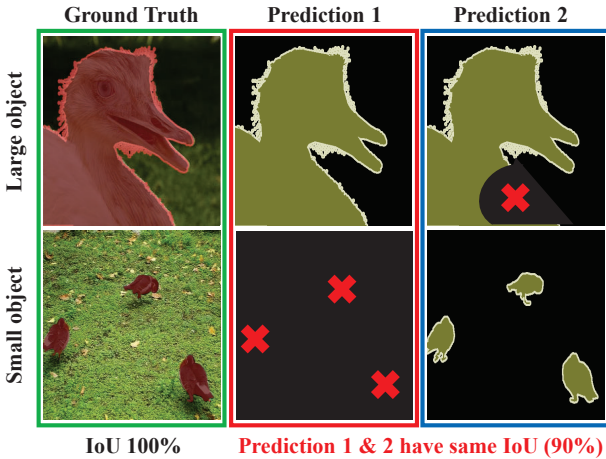(b) Imbalance problem in PASCAL VOC validation set

Figure 1. Problems of conventional metric and dataset. (a) Prediction 1 and 2 show the prediction for different cases which result in the same IoU scores. (b) Some classes of PASCAL VOC validation set suffer from a lack of small-sized objects. We sort the number of instances in descending order for each class per each size.

all distribution for 20 classes of the VOC validation set per each size[†]. Many classes fall short in the number of small objects. Even with an ideal metric, we will never know how methods perform on small objects with few samples such as small birds. Besides, we note that MS COCO [33], another popular benchmark with 80 classes for WSSS, also suffers from imbalanced distribution. More information of dataset distribution is in the supplementary material.

Training objective. Pixel-wise cross-entropy considers all individual pixels equally important by averaging. Thus the networks will consider small objects less important and lean toward large objects with many pixels. While the fully-supervised semantic segmentation methods have some remedy [12, 34], WSSS literature has paid less attention to this

---

[†]Following MS COCO, we regard an instance as small if total number of pixels$< 32 \times 32$, medium if the total number of pixels$< 96 \times 96$, and large for the rest.
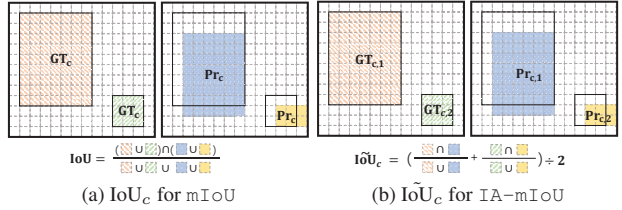

$$\text{IoU} = \frac{(\square \cup \square) \cap (\square \cup \square)}{(\square \cup \square) \cup (\square \cup \square)}$$

(a) IoU_c for `mIoU`

$$\tilde{\text{IoU}}_c = (\frac{\square \cap \square}{\square \cup \square} + \frac{\square \cap \square}{\square \cup \square}) \div 2$$

(b) $\tilde{\text{IoU}}_c$ for `IA-mIoU`

Figure 2. Visual comparison of the computing process of IoU_c for `mIoU` and $\tilde{\text{IoU}}_c$ for `IA-mIoU` regarding a class c

problem. Existing works mostly focus on producing better pseudo masks to train the main segmentation network with the same pixel-wise cross-entropy.

Our solutions. In this paper, we suggest a way to address the above three limitations. First, we introduce a new evaluation metric for semantic segmentation, instance-aware mean intersection-over-union (`IA-mIoU`). It is important to accurately capture objects of all sizes to improve `IA-mIoU`. Next, we propose an evaluation dataset balanced in terms of object-size, PASCAL-B, which contains almost the same number of instances for each size, namely, large, medium, and small. With our new benchmark and evaluation metric, we can correctly measure the performances of existing WSSS models in terms of object size. Specifically, we re-evaluate ten state-of-the-art methods [1,26,29,30,35,38,42,43,45,50] and observe interesting results; all evaluated methods struggle in capturing small objects. Lastly, we propose a new loss function paired with a training strategy for segmentation models to balance the objective. Thorough experiments on three datasets demonstrate that our method achieves comprehensive performance boost on ten existing WSSS methods. We believe that it will serve as a strong baseline to start with toward more comprehensive performance. The code and the dataset will be publicly available for research community.

2. Instance-aware mIoU

In this section, we explain how our metric addresses the limitations of `mIoU`. Then, we compare `mIoU` and our instance-aware mIoU (`IA-mIoU`) with the results of several corner cases.

2.1. Definition of `IA-mIoU`.

In Fig. 2 (a), we visualize the way of calculating IoU_c of a class c, for `mIoU`. First, IoU_c unions all pixels of ground-truth (GT_c) and prediction (Pr_c) respectively, and then calculates the intersection of them. During the process, it does not consider which instance each pixel belongs to. As a result, `mIoU` inherently does not provide a detailed sense of performance but provides coarse judgment.

To reflect the different importance of pixels, we suggest measuring the performance of each instance individ-

(a) Case 1. One prediction segment covers one GT segment.

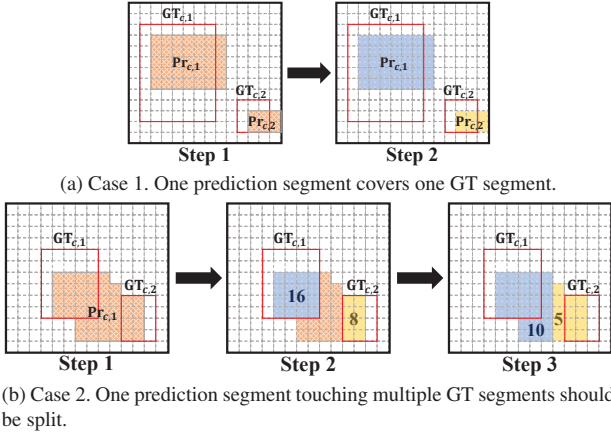(b) Case 2. One prediction segment touching multiple GT segments should be split.

Figure 3. Two cases for assigning predictions to the corresponding ground-truth instances. Pixels in color are the prediction and boxes with red lines are ground-truth instances. (a) When there is a one-to-one correspondence between prediction and ground-truth instance, each prediction is assigned to the corresponding ground-truth instance. (b) When there is a one-to-many correspondence between prediction and ground-truth instances, non-overlapping regions in step 2 (orange pixels with check pattern) distribute to each instance based on the ratio of blue and yellow pixels with dot pattern.

ually. We first split predictions and ground-truths of class c into different instances $i.e.$, $Pr_{c,1}$, $Pr_{c,2}$, $GT_{c,1}$ and $GT_{c,2}$ as shown in Fig. 2 (b). Then, we compute IoU scores $IoU_{c,i}$ for each instance i and average them to obtain $I\tilde{o}U_c$ that is instance-aware IoU score of the class c:

$$IoU_{c,i} = \frac{Pr_{c,i} \cap GT_{c,i}}{Pr_{c,i} \cup GT_{c,i}}, \quad I\tilde{o}U_c = \frac{\sum_{i=0}^{T} IoU_{c,i}}{T}, \quad (1)$$

where T is the total number of instances of the class c. Finally, we average the per-instance IoUs to compute instance-aware mIoU (IA-mIoU):

$$IA\text{-}mIoU = \frac{\sum_{c=0}^{N} I\tilde{o}U_c}{N}. \quad (2)$$

The following subsection describes how to split the predictions and ground-truths, and how to assign prediction instances to ground-truth instances.

2.2. Splitting and assigning instances

Although we introduced the concept of instance, it does not exist in the segmentation task. Hence, we assume that the ground-truth segmentation masks can be either split into connected components (blobs) or split by additional instance annotation when available for evaluation. Please note that we introduce the instance labels only for more precise evaluation, not for training.

To fully utilize the instance masks for evaluation, we also have to split the predicted segments into blobs and assign them to overlapping ground-truth instances. There are three types of predictions for the model: 1) one prediction covers one object, 2) one prediction covers multiple objects simultaneously, and 3) prediction fails to cover any target instances. We consider only the first two cases because the last case has no overlapping region between prediction and ground-truth[†]

The procedure is illustrated in Fig. 3. Both cases start from drawing contour lines from prediction for class c (Pr_c) to get connected components ($Pr_{c,i}$). The next step, however, is different for $case\ 1$ and $case\ 2$ since the former is a one-to-one correspondence relationship between $Pr_{c,i}$ and $GT_{c,i}$ and the latter is one-to-many.

For the $case\ 1$, each connected component is assigned to overlapping target instance in the second step ($Pr_{c,1} \rightarrow GT_{c,1}$ and $Pr_{c,2} \rightarrow GT_{c,2}$). Then, we can calculate the IoU per instance. On the other hand, for the $case\ 2$, we have to split the connected component into multiple parts since it overlaps with multiple target instances. In other words, we have to distribute the non-overlapped area to each instances. To do this, we apply weighted clustering algorithm that if cluster ($i.e.$, target instance) has more overlapped pixels than others, it takes larger unassigned regions. It has following advantages: 1) it does not favor or damage particular instances, 2) it is invariant to locations of the chosen pixels, and 3) it is less bias on the object size.

This algorithm is implemented by adding two steps. We first assign the intersecting regions to the corresponding target instances and compute the ratio of the overlapping area ($i.e.$, $GT_{c,1} \cap Pr_{c,1} : GT_{c,2} \cap Pr_{c,1} = 16 : 8$) in the second step. In the final step, we distribute the remaining unassigned area to each target instance according to the ratio. The way of distribution of pixels can be not unique, but we focus on reasonable distributions of pixels based on instance size. For the multiple predictions and ground-truths, we would perform the same assignment process for each prediction and its corresponding ground-truth instances. This approach enables instance-aware metric in semantic segmentation tasks, even when the model does not provide instance-level predictions. In the next subsection, we design corner cases to compare the tendencies of mIoU and IA-mIoU clearly.

2.3. Sensitivity analysis on corner cases.

We design four corner cases in Fig. 4. We first set up small and large instances in an image, and then gradually expand the predictions to cover the assigned ground-truth instances. The outcomes show the limitation of mIoU more clearly: the prediction on a large object dominates the over-

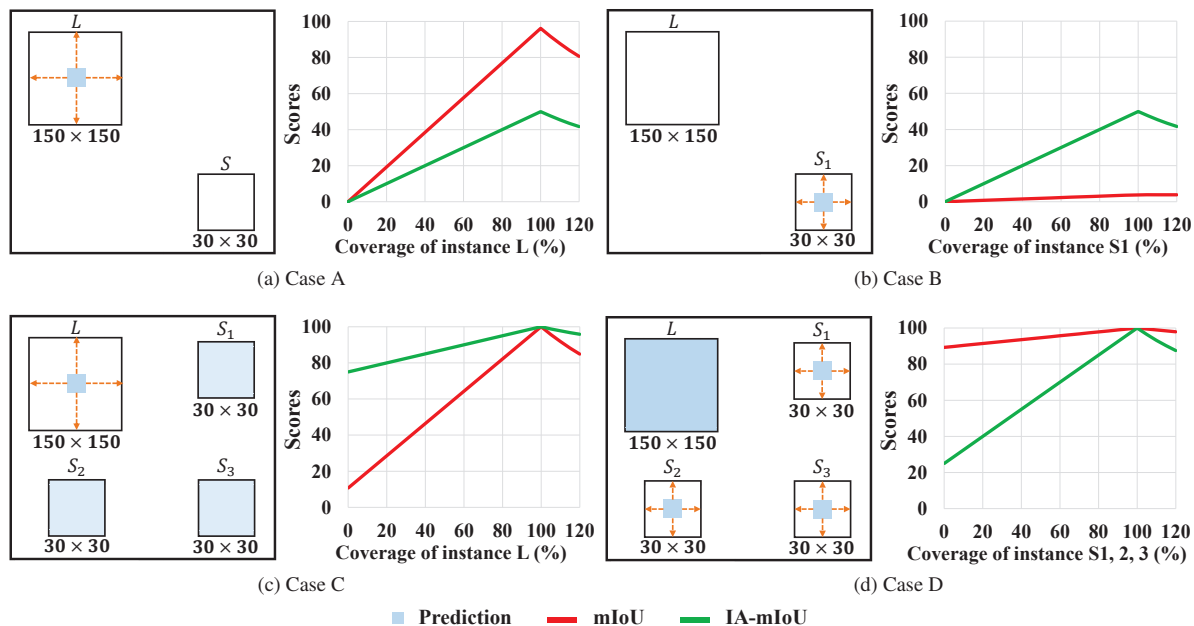[†]False positives in a class c do not contribute to $I\tilde{o}U_c$ but they decrease $I\tilde{o}U_{background}$.

Figure 4. Sensitivity to the size of instances on corner cases. We plot the behavior of mIoU and IA-mIoU as the prediction gradually grow to fill the ground-truth instance L (or $S_{1,2,3}$). Empty squares are uncovered ground-truth instances and sky blue squares are predictions. Gradual increase of the predictions is marked with orange dashed arrows.
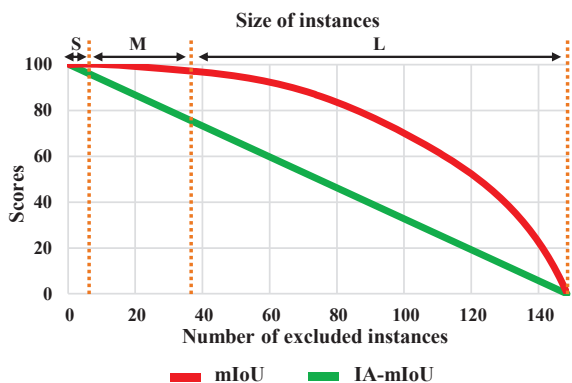

Figure 5. Corner case with real data. mIoU declines quickly as the size of instances gets larger while IA-mIoU drops consistently.

all performance. The mIoU scores of *case A* and *C* increase exponentially with the improvement of prediction on a large object. On the contrary, the performances for *case B* and *D* barely change even though the predictions on small objects improve. Unlike the mIoU, our metric IA-mIoU steadily increases as the predictions fill the target instances regardless of the instance size. Furthermore, since we split the instances, we acquire more detailed sense of the performance according to their sizes (*i.e.*, measuring only specific size of objects).

In addition, Fig. 5 plots the behavior of mIoU and

IA-mIoU in *dog* class of the PASCAL VOC 2012 dataset. Starting from the perfect score, *i.e.*, the prediction equals the ground-truth, we remove one instance at a time from the prediction starting with the smallest and progressing to the largest. IA-mIoU drops consistently, while mIoU barely decreases for small instances and rapidly decreases for large instances. We draw the red dashes in Fig. 5 to distinguish the size of instances more clearly.

We hope that it would be beneficial for the community by providing a new comprehensive evaluation metric that can measure the semantic segmentation performance on small objects accurately.

3. Dataset analysis and construction

Imbalanced evaluation dataset may cripple the reliability of an evaluation protocol because the performance will vary due to the lack of samples. We believe that any objects with various sizes should not be undervalued because of their small number.

To tackle the imbalanced dataset issue, we suggest a new balanced benchmark dataset for evaluation. We construct PASCAL-B by collecting images and annotations from LVIS [17] and MS COCO [33] datasets which includes at least one of 20 categories[†] of the PASCAL VOC classes. Then, we converted the annotations which do not belong to the 20 categories of the PASCAL VOC dataset

---

[†]From MS COCO, we only collected images of "potted-plant" since LVIS does not have it.
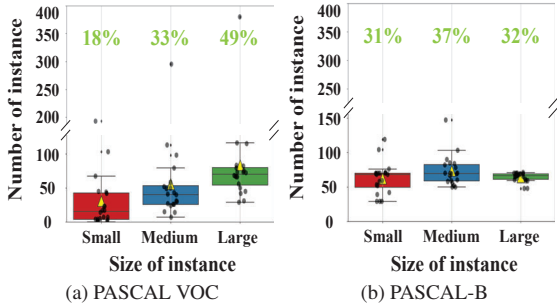
(a) PASCAL VOC (b) PASCAL-B

Figure 6. The distribution of validation set for each dataset: (a) PASCAL VOC and (b) PASCAL-B. We draw the mean (*i.e.*, the triangle in yellow) and the variance over classes for each size of instances (*i.e.*, *small*, *medium*, and *large*). The point in gray indicates the number of instances for each class. On the top of each figure, we report the ratio of each size of instances to the total number of instances.

into the background class. Among the remaining images, a few images have wrong annotations. Therefore, two computer vision experts (authors of this paper) manually filtered out such images for two weeks. Then, we randomly sampled images to ensure the balance over classes and object size distribution. In the end, PASCAL-B consists of 1,137 images with 20 classes. We give some representative images of the PASCAL-B dataset in the supplementary material.

As illustrated in Fig. 6 (b), our dataset is much more balanced in terms of classes and object-size distribution. Compared to PASCAL VOC, our PASCAL-B has fewer outliers, *i.e.*, points in gray, and they do not have extremely large values. Also, PASCAL-B keeps a similar number of instances for each size while PSACAL VOC has more large or small instances. In summary, a primary motivation for creating PASCAL-B was to address the issue of imbalanced evaluation datasets commonly encountered in semantic segmentation task. Existing benchmarks suffer from disparities in class or object size distributions, leading to skewed performance evaluations. PASCAL-B addresses this concern by meticulously constructing a dataset that features balanced classes and object sizes. Instead of replacing established benchmarks such as ADE20K [49], COCO [33], or Cityscapes [7], PASCAL-B complements them by offering an alternative approach to assessment. For more details regarding the dataset, please refer to the supplementary material.

4. Methods

4.1. Evaluated WSSS methods

We evaluate ten existing methods under various weaklevels of supervision: bounding box supervision (*i.e.*, BANA [29] and BBAM [35]), saliency supervision (*i.e.*,

Figure 7. Example connected components for the loss function. $I_{c,k}$ is the k-th connected components for c-th class in an image.

RCA [50], EDAM [42] and NS-ROM [45]), natural language supervision (*i.e.*, CLIM [43]), and image supervision (*i.e.*, AMN [30], RIB [26], CDA [38], and IRN [1]). These methods follow the two-stage training pipeline of WSSS. In the first stage, they generate the pseudo masks by their methods. Then, they train a semantic segmentation network with the pseudo masks from the first stage. All the above methods except BANA [35] only focus on stage 1 to produce the high-quality masks by refining the initial seed to improve the performance. A more detailed explanation for the above methods is in the supplementary material.

4.2. Proposed loss function and training strategy

To address the limitation of pixel-wise cross-entropy (CE) loss in Sec. 1, we propose a new loss function for a model to have the capacity of capturing small objects. We first give weights to each pixel according to the size of the object when computing the loss. Since the instance ground-truth masks are not available for training, we find all connected components for each class from pseudo ground-truth masks as in Fig. 7. Then, we get weight $w_{x,y}$ corresponding to a pixel (x, y) as follows:

$$w_{x,y} = \begin{cases} 1, & \text{if } (x,y) \in background, \\ \min(\tau, \frac{\sum_{k=1}^{K} \text{S}_{c,k}}{\text{S}_{c,n}}), & \text{otherwise} \end{cases} \quad (3)$$

where $\text{S}_{c,k}$ is the number of pixels in its connected component $I_{c,k}$, while n is the index number of instance which pixel (x, y) is included. K is the number of instances with c-th class in an image. Through Eq. 3, we assign a larger weight to the pixels of the relatively small instance while preventing the value of weight from getting excessively large by setting up the upper limit τ. Finally, we multiply weights to cross-entropy loss as in Eq. 4 and we call this loss function L_{sw} as size-weighted cross-entropy loss.

$$L_{sw} = -\frac{1}{H \times W} \sum_{c=1}^{C} \sum_{x=1}^{H} \sum_{y=1}^{W} Y_{c,x,y} w_{x,y} log(p_{c,x,y}), \quad (4)$$

where H and W is the height and width of images, respectively, and $p_{c,x,y}$ is the probability to predict the class of the pixel (x, y) as c.

Even though L_{sw} can improve the ability of the model to catch small objects, there is a side effect that the model fails to learn extremely large instances with L_{sw} during the whole training process. Therefore, we apply a new training strategy that adds a regularization term to Eq. 4 by introducing elastic weight consolidation (EWC) [11]. EWC helps model to learn new tasks continually while preserving the information of previous tasks. Following the strategy of EWC, we also divide the training into two tasks. We first train a model using pixel-wise cross-entropy loss which is more beneficial to learn the large object as we analyze in Sec. 1, and call this task as *task A*. During the training for *task A*, model updates the importance of parameters in Fisher information matrix. Then, for the new task, the model is fine-tuned by L_{sw} and EWC helps to regularize the important parameter for the previous *task A* based on the matrix. Thus, our final loss function L_{sb}, size-balanced cross-entropy loss, is defined as:

$$L_{sb} = L_{sw} + \sum_i \frac{\lambda}{2} F_i (\theta_i - \theta_{A,i}^*)^2, \qquad (5)$$

where θ_i and $\theta_{A,i}^*$ are i-th parameter for present task and *task A*, respectively. λ controls the importance of regularization and F_i is the importance of parameter i in the Fisher information matrix. With L_{sb}, a model can learn the new information for *task B* (*i.e.,* learning small objects) while maintaining the previous information from *task A* (*i.e.,* learning large objects).

5. Experiments

5.1. Experimental setting

Dataset. We evaluate each method on three datasets: PASCAL VOC [13], PASCAL-B, and MS COCO [33]. PASCAL VOC and PASCAL-B share the same training set though PASCAL-B is only designed for validation rather than training. PASCAL VOC and PASCAL-B consist of a similar number of images, 1,449 and 1,137, respectively.

Evaluation metric. We use mIoU and IA-mIoU to compare the performance of methods. Since our IA-mIoU can measure the small-sized instance only, we provide the IA_S for the detailed performance of small objects.

Implementation detail. We generate pseudo masks for the segmentation networks using the official codes and strictly follow the setting provided in each paper [1, 26, 29, 30, 35, 38, 42, 43, 45, 50]. Then, we use DeeplabV2 with ResNet-101 [4] as segmentation networks. For more detail, please see the supplementary material. All the experiments were done by one GeForce RTX 3090 GPU for PASCAL VOC and two RTX 3090 GPUs for MS COCO.

Method	Sup.	mIoU	IA-mIoU	IA_S
IRN*	$\mathcal{I}$	64.8	56.0	17.5
CDA*	$\mathcal{I}$	66.6	57.2	15.8
AMN*	$\mathcal{I}$	69.4	58.4	15.9
CLIM*	$\mathcal{I}, \mathcal{L}$	68.9	57.4	14.0
RCA†	$\mathcal{I}, \mathcal{S}$	70.4	60.7	23.2
EDAM†	$\mathcal{I}, \mathcal{S}$	70.7	60.7	21.3
NS-ROM†	$\mathcal{I}, \mathcal{S}$	70.4	60.2	19.3
BANA†	$\mathcal{I}, \mathcal{B}$	72.6	59.6	14.7
BBAM*	$\mathcal{I}, \mathcal{B}$	72.7	60.5	14.7

Table 1. Experimental results for PASCAL VOC. ∗ and † indicate that the segmentation model utilizes the ImageNet and COCO pre-trained model respectively. $\mathcal{I}, \mathcal{S}, \mathcal{L}$ and $\mathcal{B}$ denotes the degree of supervision. $\mathcal{I}$: image-level supervision, $\mathcal{L}$: natural language supervision, $\mathcal{S}$: saliency supervision, and $\mathcal{B}$: bounding box supervision.

Method	Sup.	mIoU	IA-mIoU	IA_S
DeepLabV2*	$\mathcal{F}$	55.4	33.5	12.9
RIB*	$\mathcal{I}$	44.6	29.2	11.4
IRN*	$\mathcal{I}$	39.7	25.8	9.4

Table 2. Experimental results for MS COCO.

5.2. Quantitative results

We evaluate nine baseline methods on PASCAL VOC and PASCAL-B, and three baseline methods on MS COCO. Furthermore, we demonstrate that our size-balanced cross-entropy loss function on the baseline methods results in better segmentation performance when compared to using the conventional cross-entropy (CE) loss.

mIoU vs. IA-mIoU. Table 1 compares the performances in mIoU and IA-mIoU on the PASCAL VOC dataset. Although the recent WSSS methods make impressive performance in the mIoU metric, we observe that the detailed scores measured by IA-mIoU are quite different. It is noteworthy that all WSSS methods get badly lower scores for small objects (IA_S) compared to overall scores. It indicates that WSSS methods struggle to capture the small instances accurately as we mentioned in Sec. 1.

In particular, state-of-the-art techniques in terms of mIoU encounter more difficulty in capturing small objects compared to other methods. Consequently, they get lower IA-mIoU while getting the highest mIoU, since IA-mIoU reflects the scores of each instance equally but mIoU relatively neglects the small objects. This indicates mIoU fails to catch the detailed sense of performance on different sizes of objects.

We do the same experiments on the MS COCO dataset in Table 2. According to the result of these experiments, we further demonstrate that existing WSSS methods struggle with small objects and it has been overlooked with mIoU.

Method	Sup.	mIoU	IA-mIoU	IA$_S$
IRN*	$\mathcal{I}$	56.1	41.0	15.8
CDA*	$\mathcal{I}$	57.5	41.4	13.4
AMN*	$\mathcal{I}$	58.5	41.1	13.9
CLIM*	$\mathcal{I,L}$	58.7	40.2	12.2
RCA*	$\mathcal{I,S}$	60.8	45.5	18.4
EDAM†	$\mathcal{I,S}$	60.4	45.2	19.4
NS-ROM†	$\mathcal{I,S}$	58.9	43.6	16.2
BANA†	$\mathcal{I,B}$	61.9	41.1	14.0
BBAM*	$\mathcal{I,B}$	60.1	40.9	14.3

Table 3. Experimental results for PASCAL-B.

PASCAL VOC vs. PASCAL-B. Table 3 compares the performances of models on our newly proposed benchmark, PASCAL-B. The models in Table 3 use the same checkpoint from Table 1 which are trained using the PASCAL VOC training set.

We argue that evaluating methods using imbalanced datasets can lead to biased scores, even with our proposed metric. To better evaluate the ability of models, it is essential to have a sufficient number of samples for evaluation per object-size and per class. However, the imbalance in the PASCAL VOC dataset makes it difficult to validate models since some classes have no small-sized objects, or there are only a few samples available. This lack of data for certain classes limits the opportunities for models to be evaluated on their performance, leading to potential biases in the evaluation process. On the other hand, we address this issue by constructing PASCAL-B which includes a sufficient number of samples for each object-size while keeping a balanced distribution across classes.

In this manner, the results in Table 3 with PASCAL-B provide better comprehensive assessment of WSSS methods compared to the scores in Table 1. When comparing the results of both tables, we observe that ranking order of WSSS methods is barely changed for mIoU and IA-mIoU in Table 1 (Spearman's rho: 0.79). On the other hand, it has totally changed in Table 3 with PASCAL-B dataset (Spearman's rho: 0.38), which indicates that IA-mIoU scores with PASCAL-B evaluates the performance of models differently. We believe that the fundamental reason for this phenomenon lies in the discrepancy of distributions in terms of instance sizes between the two datasets. This suggests that IA-mIoU and PASCAL-B are both necessary to properly evaluate per-size performances.

CE loss vs. Size-balanced CE loss. Lastly, we verify the effectiveness of our proposed method, size-balanced cross-entropy loss function. As shown in Fig. 8, our method successfully boosts the performances for all models across three datasets. In particular, it enhances the ability of models to capture small instances. Across all datasets, we ob-

(a) PASCAL VOC (IA-mIoU) (b) PASCAL VOC (IA$_S$)

(c) PASCAL B (IA-mIoU) (d) PASCAL B (IA$_S$)

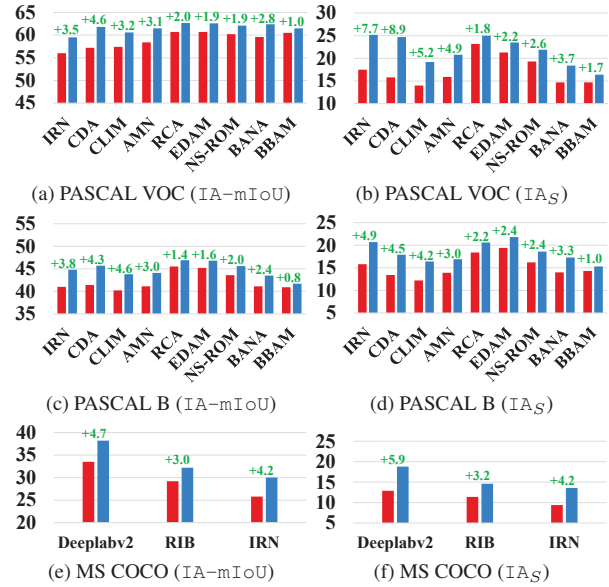(e) MS COCO (IA-mIoU) (f) MS COCO (IA$_S$)

Figure 8. Comparison of experimental results when applying CE loss (red bar) and Size-balanced CE loss (blue bar). We mark the increment above the bar (number with green color.)

serve an increase of IA$_S$ scores ranging from 1.0 to 8.9. The changes of mIoU values, however, are relatively negligible, since the increase in performance of catching small instances has a little impact on mIoU as we explained in Sec. 2 (Out of 21 experiments, 18 have shown slight improvements in mIoU). In the supplementary material, we analyze qualitatively the experimental results according to the usage of our proposed loss function and provide more detailed values of performance gain.

5.3. Ablation study

In this subsection, we demonstrate the effectiveness of each component of our loss function on the PASCAL VOC dataset with mIoU and IA-mIoU. In Table 4, we use a fully-supervised method, DeepLabV2 [4] as our baseline model to observe performance gains by adding our components to the baseline.

Method	mIoU	IA-mIoU	IA$_S$
DeepLabV2	77.8	65.8	18.8
with L_{sw}	77.5	68.7	23.0
with L_{sb}	78.4	69.5	24.4

Table 4. Ablation study on each component of our loss function. L_{sb}: Add regularization to L_{sw} using EWC.

Applying only the size-weighted cross-entropy loss function L_{sw} is powerful enough to gain notable improvements on small instances (IA$_S$) and IA-mIoU increases by 2.9 points. However, mIoU becomes slightly worse than

the baseline. In other words, L_{sw} alone does not ensure the same performance on the largest instances. L_{sb} further boosts performance in all aspects by facilitating additional objective, covering small instances, while maintaining the previous objective, covering relatively large instances. Again, `IA-mIoU` enables detailed analyses by splitting the instances. In short, introducing the size-balanced cross-entropy loss improves the performance on small instances and pairing EWC training strategy preserves the performance on large instances, resulting in overall improvement in both `mIoU` and `IA-mIoU`.

6. Related Work

6.1. Weakly-supervised semantic segmentation

Weakly-supervised semantic segmentation mainly adopts a two-stage pipeline: pseudo mask generation and training segmentation network. Most recent methods utilize Class Activation Maps (CAMs) [48] to generate a pseudo mask. However, CAMs have limitations in focusing on the most discriminative regions of the object or capturing frequently co-occurring background components. To solve this problem, lots of techniques have been proposed: adversarial erasing [6, 18, 25, 32, 40, 41], seed growing [19, 23, 46], natural language supervision [43], context decoupling [38] and so on [2, 3, 26, 28, 47]. Also, many methods [14–16, 20, 27, 39, 42, 44, 45] adopt a saliency supervision to refine the prediction map. It is usually utilized to enhance the result in a post-processing step by distinguishing the foreground and background of the object. Recently, Lee et al. [31] try to make use of a saliency map during the training phase to maximize its potential. Besides, there are also some studies using a bounding box as a supervisory signal [10, 21, 24, 29, 35–37] which is still cheaper than mask annotation. They achieve notable performance in WSSS since a bounding box label provides the exact location of all objects additionally. Our research, however, is interested in getting the better performance of models by improving the segmentation network in the second stage. Though few studies propose methods for segmentation networks, we suggest balanced training considering the size of instances in WSSS.

6.2. Segmentation metrics

Here we briefly review the metrics for semantic segmentation. Pixel accuracy is the most basic metric for the task. It measures the accuracy for each class by computing the ratio of correctly predicted pixels of the class to all pixels of that class. The weakness of this metric is it does not consider false positives. Therefore, mean intersection-over-union (`mIoU`) replaces the pixel accuracy for semantic segmentation measures. It assesses the performance of models by calculating prediction masks intersection ground-truth masks over prediction masks union ground-truth masks. The `mIoU` compensates for the shortcoming of pixel accuracy by taking account of false positive. Nonetheless, as we analyze it in the next section, it still suffers from a size imbalance problem. Besides, various metrics [5, 8, 9, 22] are also investigated. Cordts et al. [8] point out the inherent bias of the traditional IoU measure towards larger instances. They proposed instance-level IoU which focuses on adjusting pixel contributions based on instance sizes and class-averaged instance sizes, aiming to refine mIoU. However, our metric IA-mIoU evaluates each instance individually by segmenting predictions into instances, providing a comprehensive assessment that is not influenced by instance size.

7. Conclusion

7.1. Contributions

In this paper, we focus on the comprehensive assessment and improvement of weakly-supervised semantic segmentation (WSSS) by proposing a novel metric, dataset, and loss function with an appropriate training strategy. First, we uncover the overlooked issue related to small-sized instances due to the conventional metric (`mIoU`). To address this, we design the instance-aware mIoU (`IA-mIoU`) to measure the performance of models more precisely regardless of object-size. Moreover, we point out the imbalance problem in benchmarks of WSSS and introduce a well-balanced dataset for evaluation, PASCAL-B. Lastly, we propose the size-balanced cross-entropy loss to compensate for the imbalance problem of pixel-wise cross-entropy loss. We show the effectiveness of our loss function on ten WSSS methods over three datasets measured by `mIoU` and `IA-mIoU`.

7.2. Limitations

Our findings can be applied to fully-supervised semantic segmentation methods. However, due to limited computing power, we were unable to utilize more recent FSSS models and evaluate them with datasets such as ADE20K [49] and Cityscapes [7]. Nevertheless, we hope that our study can serve as inspiration for other researchers who have the necessary resources to explore these avenues further.

7.3. Acknowledgment

References

[1] Jiwoon Ahn, Sunghyun Cho, and Suha Kwak. Weakly supervised learning of instance segmentation with inter-pixel relations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2209–2218, 2019. 1, 2, 5, 6

[2] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 8991–9000, 2020. 8

[3] Liyi Chen, Weiwei Wu, Chenchen Fu, Xiao Han, and Yuntao Zhang. Weakly supervised semantic segmentation with boundary exploration. In *European Conference on Computer Vision*, pages 347–362. Springer, 2020. 8

[4] 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. 1, 6, 7

[5] Bowen Cheng, Ross Girshick, Piotr Dollár, Alexander C Berg, and Alexander Kirillov. Boundary iou: Improving object-centric image segmentation evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15334–15342, 2021. 8

[6] 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*, 43(12):4256–4271, 2020. 8

[7] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, pages 3213–3223, 2016. 5, 8

[8] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 8

[9] Gabriela Csurka, Diane Larlus, Florent Perronnin, and France Meylan. What is a good evaluation measure for semantic segmentation? *IEEE PAMI*, 26(1), 2004. 8

[10] Jifeng Dai, Kaiming He, and Jian Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *ICCV*, pages 1635–1643, 2015. 8

[11] Kirkpatrick et al. Overcoming catastrophic forgetting in neural networks. *PNAS*, 2017. 6

[12] Lin et al. Focal loss for dense object detection. In *ICCV*, 2017. 2

[13] 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. 1, 6

[14] Junsong Fan, Zhaoxiang Zhang, Chunfeng Song, and Tieniu Tan. Learning integral objects with intra-class discriminator for weakly-supervised semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 8

[15] Junsong Fan, Zhaoxiang Zhang, and Tieniu Tan. Employing multi-estimations for weakly-supervised semantic segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*, pages 332–348. Springer, 2020. 8

[16] 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. 8

[17] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5356–5364, 2019. 4

[18] Qibin Hou, PengTao Jiang, Yunchao Wei, and Ming-Ming Cheng. Self-erasing network for integral object attention. In *NeurIPS*, pages 547–557, 2018. 8

[19] 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. 8

[20] 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/CVF International Conference on Computer Vision*, pages 2070–2079, 2019. 8

[21] Anna Khoreva, Rodrigo Benenson, Jan Hosang, Matthias Hein, and Bernt Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In *CVPR*, pages 876–885, 2017. 8

[22] Pushmeet Kohli, Philip HS Torr, et al. Robust higher order potentials for enforcing label consistency. *International Journal of Computer Vision*, 82(3):302–324, 2009. 8

[23] Alexander Kolesnikov and Christoph H Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *European conference on computer vision*, pages 695–711. Springer, 2016. 8

[24] Viveka Kulharia, Siddhartha Chandra, Amit Agrawal, Philip Torr, and Ambrish Tyagi. Box2seg: Attention weighted loss and discriminative feature learning for weakly supervised segmentation. In *European Conference on Computer Vision*, pages 290–308. Springer, 2020. 8

[25] Hyeokjun Kweon, Sung-Hoon Yoon, Hyeonseong Kim, Daehee Park, and Kuk-Jin Yoon. Unlocking the potential of ordinary classifier: Class-specific adversarial erasing framework for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6994–7003, 2021. 8

[26] Jungbeom Lee, Jooyoung Choi, Jisoo Mok, and Sungroh Yoon. Reducing information bottleneck for weakly supervised semantic segmentation. *Advances in Neural Information Processing Systems*, 34, 2021. 1, 2, 5, 6, 8

[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/CVF Conference on Computer Vision and Pattern Recognition*, pages 5267–5276, 2019. 8

[28] Jungbeom Lee, Eunji Kim, and Sungroh Yoon. Anti-adversarially manipulated attributions for weakly and semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4071–4080, 2021. 8

[29] 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. 1, 2, 5, 6, 8

[30] Minhyun Lee, Dongseob Kim, and Hyunjung Shim. Threshold matters in wsss: Manipulating the activation for the robust and accurate segmentation model against thresholds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4330–4339, 2022. 1, 2, 5, 6

[31] 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. 8

[32] 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. 8

[33] 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 *ECCV*, pages 740–755, 2014. 2, 4, 5, 6

[34] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015. 2

[35] Youngmin Oh, Beomjun Kim, and Bumsub Ham. Background-aware pooling and noise-aware loss for weakly-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6913–6922, 2021. 1, 2, 5, 6, 8

[36] George Papandreou, Liang-Chieh Chen, Kevin P Murphy, and Alan L Yuille. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *ICCV*, pages 1742–1750, 2015. 8

[37] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. Box-driven class-wise region masking and filling rate guided loss for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3136–3145, 2019. 8

[38] Yukun Su, Ruizhou Sun, Guosheng Lin, and Qingyao Wu. Context decoupling augmentation for weakly supervised semantic segmentation. *arXiv preprint arXiv:2103.01795*, 2021. 1, 2, 5, 6, 8

[39] Guolei Sun, Wenguan Wang, Jifeng Dai, and Luc Van Gool. Mining cross-image semantics for weakly supervised seman-tic segmentation. In *European conference on computer vision*, pages 347–365. Springer, 2020. 8

[40] Kunyang Sun, Haoqing Shi, Zhengming Zhang, and Yong-ming Huang. Ecs-net: Improving weakly supervised semantic segmentation by using connections between class activation maps. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7283–7292, 2021. 8

[41] 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. 8

[42] Tong Wu, Junshi Huang, Guangyu Gao, Xiaoming Wei, Xi-aolin Wei, Xuan Luo, and Chi Harold Liu. Embedded discriminative attention mechanism for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16765–16774, 2021. 1, 2, 5, 6, 8

[43] Jinheng Xie, Xianxu Hou, Kai Ye, and Linlin Shen. Clims: Cross language image matching for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4483–4492, June 2022. 1, 2, 5, 6, 8

[44] Lian Xu, Wanli Ouyang, Mohammed Bennamoun, Farid Boussaid, Ferdous Sohel, and Dan Xu. Leveraging auxiliary tasks with affinity learning for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6984–6993, 2021. 8

[45] Yazhou Yao, Tao Chen, Guo-Sen Xie, Chuanyi Zhang, Fumin Shen, Qi Wu, Zhenmin Tang, and Jian Zhang. Non-salient region object mining for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2623–2632, 2021. 1, 2, 5, 6, 8

[46] 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. 8

[47] Dong Zhang, Hanwang Zhang, Jinhui Tang, Xiansheng Hua, and Qianru Sun. Causal intervention for weakly-supervised semantic segmentation. *arXiv preprint arXiv:2009.12547*, 2020. 8

[48] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, pages 2921–2929, 2016. 8

[49] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5122–5130, 2017. 5, 8

[50] Tianfei Zhou, Meijie Zhang, Fang Zhao, and Jianwu Li. Regional semantic contrast and aggregation for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4299–4309, 2022. 1, 2, 5, 6