

BBAM: Bounding Box Attribution Map for Weakly Supervised Semantic and Instance Segmentation

Jungbeom Lee¹Jihun Yi¹Chaehun Shin¹Sungroh Yoon^{1,2,*}¹ Department of Electrical and Computer Engineering, Seoul National University, Seoul, South Korea² ASRI, INMC, ISRC, and Institute of Engineering Research, Seoul National University

{jbeom.lee93, t080205, chaehuny, sryoon}@snu.ac.kr

Abstract

Weakly supervised segmentation methods using bounding box annotations focus on obtaining a pixel-level mask from each box containing an object. Existing methods typically depend on a class-agnostic mask generator, which operates on the low-level information intrinsic to an image. In this work, we utilize higher-level information from the behavior of a trained object detector, by seeking the smallest areas of the image from which the object detector produces almost the same result as it does from the whole image. These areas constitute a bounding-box attribution map (BBAM), which identifies the target object in its bounding box and thus serves as pseudo ground-truth for weakly supervised semantic and instance segmentation. This approach significantly outperforms recent comparable techniques on both the PASCAL VOC and MS COCO benchmarks in weakly supervised semantic and instance segmentation. In addition, we provide a detailed analysis of our method, offering deeper insight into the behavior of the BBAM. The code is available at: <https://github.com/jbeomlee93/BBAM>.

1. Introduction

Object segmentation is one of the most important steps in image recognition. Advances in deep learning have greatly improved the performance of semantic and instance segmentation [8, 23] through the use of huge amounts of pixel-level annotated training data. However, annotating with pixel-level masks requires a lot of effort. According to Bearman *et al.* [4], constructing a pixel-level mask for an image containing an average of 2.8 objects takes about 4 minutes. This is why weakly supervised methods have been proposed, in which segmentation networks are trained using annotations that are less detailed than pixel-level masks, such as bounding boxes [11, 31, 60], or image-level tags [1, 2, 36].

The most easily obtainable annotation is the class label.

Labeling an image with class labels takes around 20 seconds [4], but it only indicates that objects of certain classes are depicted and gives no information about their locations in the image. Moreover, class labels provide no help in separating different objects of the same class, which is the goal of instance segmentation.

Bounding boxes provide information about individual objects and their locations. Bounding box annotation takes about 38.1 seconds per image [5], which is much more attractive than constructing pixel-level masks. Many researchers have tackled semantic segmentation [11, 31, 34, 60] and instance segmentation [3, 27, 31, 40, 62] using bounding box annotations as a search space in which a class-agnostic object mask can be found by an off-the-shelf object mask generator. These are mostly based on GrabCut [53] or multiscale combinatorial grouping (MCG) [49]. Those mask generators operate on the low-level information of images, such as the color or brightness of pixels, and this limits the quality of the resulting mask. Thus, applying these mask generators to bounding box annotations requires additional steps such as estimating what proportion of the pixels in a bounding-box belong to the corresponding object [34, 60], iterative refinement of an estimated mask [11], and auxiliary attention modules [34].

We propose a pixel-level method of localizing a target object inside its bounding box using a trained object detector. We make use of attribution maps obtained from the trained object detector, which highlight the image regions that the detector focuses on in conducting object detection. Inspired by the perturbation methods used to explain the output of image classifiers [10, 17, 18], we introduce a bounding box attribution map (BBAM) which provides an indication of the smallest areas of an image that are sufficient to make an object detector produce almost the same result as that from the original image. The BBAM identifies the area occupied by the object in each bounding box predicted by the trained object detector. Since this localization takes place at the pixel level, it can be used as a pseudo ground truth for weakly supervised learning of semantic and instance segmentation.

*Correspondence to: Sungroh Yoon <sryoon@snu.ac.kr>.

The main contributions of this paper can be summarized as follows.

- We propose a bounding box attribution map (BBAM), which can draw on the rich semantics learned by an object detector to produce pseudo ground-truth for training semantic and instance segmentation networks.
- Our technique significantly outperforms previous state-of-the-art methods of weakly supervised semantic and instance segmentation, assessed on the PASCAL VOC 2012 and MS COCO 2017 benchmarks.
- We analyze our method from various viewpoints, providing deeper insights into the properties of the BBAM.

2. Related Work

Fully supervised semantic and instance segmentation based on pixel-level annotations is highly reliable, but the manual annotation process is laborious. This requirement is overcome by weakly supervised methods based on inexact, but easily obtainable, annotations such as scribbles [63], bounding boxes [31, 60], or class labels [1, 36, 61]. In this section, we briefly review some recently introduced weakly supervised approaches that use class labels (Section 2.1) or bounding boxes (Section 2.2). In addition, we describe some visual saliency methods related to our method (Section 2.3).

2.1. Learning with Class Labels

A class activation map (CAM) [69] is a widely adopted technique to obtain a localization map from class labels. However, a CAM only identifies the most discriminative regions of objects [36, 37], and hence the majority of existing methods that use class labels [2, 15, 24, 25, 28, 30, 36, 37, 38, 39, 58] are primarily concerned with expanding the area of the target object activated by a CAM. For instance, erasure methods [25, 64] iteratively find new regions of the target object by removing discriminative regions in an image. Other methods [15, 61] consider the information shared between several images by capturing cross-image semantic similarities and differences. Seed growing and refinement techniques [1, 2, 28] are typically used to expand the regions representing the target object imperfectly that are in the initial CAM, on the basis of relationships between pixels. Other methods construct CAMs that embody the multi-scale semantic context in an image [36, 38, 65]. Despite these efforts, the information available from class labels remains limited, so auxiliary information acquired from web images [56] or videos [24, 37] can be used together.

2.2. Learning with Bounding Boxes

Class labels have led to significant achievements in semantic segmentation, but they are inherently unhelpful in instance segmentation, which requires the separation of different objects of the same class. In contrast, bounding boxes

do provide information about the location of individual objects in an image, and they are still much cheaper than constructing pixel-level masks [5]. Most existing methods utilized a bounding box as a search space to conduct low-level searches for object masks. They create a pseudo mask within a box using off-the-shelf methods of mask proposal such as MCG [49] or GrabCut [53]. These processes can be guided by specifying the proportion of the pixels in a bounding box that are likely to belong to the object [34, 60]. Iterative mask refinement techniques [11] can also be applied. However, these methods are largely based on low-level information in the image, and they ignore the semantics associated with the bounding boxes. A rare exception is the multiple-instance learning formulation with a bounding box tightness prior [27]: a crossing line within a box must contain at least one pixel of the target object. The drawback with this approach is that only a small number of pixels are contributing to the localization of the object.

2.3. Visual Saliency Methods

Various methods have been proposed to visually explain the predictions of deep neural networks (DNNs) [6, 17, 18, 54, 69] in a form of a saliency map. However, most studies have been concerned with classifiers, and only a few have looked at DNNs performing other tasks [26, 51]. In particular, there have been no attempts to explain the predictions of object detectors, except Wu *et al.* [66], who embedded interpretability inside the DNN, in this case Faster R-CNN [52]. However, the explanation produced by their modified DNN is not immediately understandable because it is given as a form of tree, and thus it is not appropriate to generate pseudo ground truth for weakly supervised segmentation. Gradient-based methods, such as SimpleGrad [68], SmoothGrad [59], and Grad-CAM [55], can provide visual saliency maps of the results from classifiers, but these methods are not easily extended to object detectors, because of the structural difference between classifiers and object detectors. Nevertheless, gradient-based methods have a significant bearing on our approach, and we look at them in more detail in Section 5.

3. Method

We first provide a brief description of the operation of object detectors in Section 3.1. In Section 3.2, we introduce the BBAM for localizing objects in the bounding box. We then utilize the BBAM for weakly supervised semantic and instance segmentation in Sections 3.3 and 3.4.

3.1. Revisiting Object Detectors

Modern object detectors can be divided into two categories: one-stage [41, 43, 50] and two-stage [20, 52] approaches. We focus on two-stage object detectors such as Faster R-CNN [52], in which the two stages are region proposal and box refinement. A region proposal network (RPN)

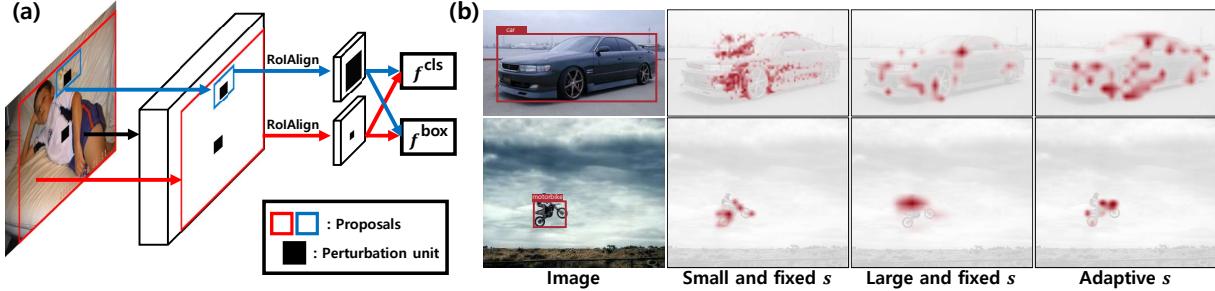


Figure 1: The size of the *perturbation unit* needs to be adjusted to the object size. (a) RoIAlign [23] produces perturbation units of different sizes. (b) Examples of resulting BBAMs with small fixed values of s , large fixed values of s , and values of s determined adaptively. Fixed values of s , whether large or small, tend to generate unwanted artifacts.

generates candidate object proposals in the form of bounding boxes; but these proposals are class-agnostic and noisy, and most of them are redundant, thereby necessitating a subsequent refinement step, in which classification and bounding box regression are performed on each proposal. Since the proposal boxes proposed by the RPN are of different sizes, RoI pooling (*e.g.*, RoIAlign [23]) is used to convert the feature map corresponding to each proposal to a predefined fixed size, as shown in Figure 1(a). The pooled feature map is then passed to the *classification head* and also to the *bounding box regression head*.

Classification head. It computes the class probability p^c of class c for each proposal and assigns the most likely class $c^* = \text{argmax}_c p^c$ to the proposal.

Bounding box regression head. It adjusts the noisy proposal to fit the object by computing the offsets $t^c = (t_x^c, t_y^c, t_w^c, t_h^c)$ for each class $c \in \{1, 2, \dots, C\}$. The final localization is obtained by shifting each coordinate of the proposal using the offset t^{c^*} . We refer to Ren *et al.* [52] for the details of the parameterization of each coordinate.

For simplicity, we will abbreviate *classification head* and *bounding box regression head* as *cls head* and *box head*, respectively.

3.2. Bounding Box Attribution Map

Suppose we are given an image I and the corresponding bounding box annotations. We also have a set of object proposals $\mathcal{O} = \{o_k\}_{k=1}^K$, either given or obtained by RPN, where K is the number of proposals. For each proposal o_k , the *box head* f^{box} and the *cls head* f^{cls} produce box offsets $t_k = f^{\text{box}}(I, o_k)$ and the class probability $p_k = f^{\text{cls}}(I, o_k)$, respectively. We omit the proposal indices k for brevity.

The bounding box attribution map (BBAM) identifies the important region in the image that the detector needs to perform object detection. We find the smallest mask $\mathcal{M} : \Omega \rightarrow [0, 1]$ where Ω is a set of pixels, which captures a subset of the image that produces almost the same prediction as the original image. A small \mathcal{M} reduces the amount of unnecessary information reaching the detector. The mask specifies a subset of the image in terms of the perturbation

function $\Phi(I, \mathcal{M}) = I \circ \mathcal{M} + \mu \circ (1 - \mathcal{M})$, where \circ denotes pixel-wise multiplication, and μ is the per-channel mean of the training data with the same size as \mathcal{M} . For each proposal o , the best mask \mathcal{M}^* is obtained by optimizing the following function using gradient descent with respect to \mathcal{M} :

$$\mathcal{M}^* = \underset{\mathcal{M} \in [0, 1]^{\Omega}}{\operatorname{argmin}} \lambda \|\mathcal{M}\|_1 + \mathcal{L}_{\text{perturb}}, \quad (1)$$

$$\begin{aligned} \mathcal{L}_{\text{perturb}} = & \mathbb{1}_{\text{box}} \|t^c - f^{\text{box}}(\Phi(I, \mathcal{M}), o)\|_1 \\ & + \mathbb{1}_{\text{cls}} \|p^c - f^{\text{cls}}(\Phi(I, \mathcal{M}), o)\|_1, \end{aligned} \quad (2)$$

where $\mathbb{1}_{\text{box}}$ and $\mathbb{1}_{\text{cls}}$ are logical variables that have a value of 0 or 1, to control which head is used to produce localizations, and $t^c = f^{\text{box}}(I, o)$ and $p^c = f^{\text{cls}}(I, o)$ are the predictions for the original image.

Previous studies show that using a mask of the same spatial size as the input image incurs undesirable artifacts due to the adversarial effect [21]: even a perturbation in a tiny magnitude can significantly change the prediction of a DNN. This problem can be addressed by introducing a coarse mask downsampled by a stride s [10, 17, 18, 26], so multiple image pixels are perturbed by a single element of \mathcal{M} . We can then optimize $\mathcal{M} \in \mathbb{R}^{\lceil w/s \rceil \times \lceil h/s \rceil}$ for the image $I \in \mathbb{R}^{w \times h}$, using the perturbation function $\Phi(I, \mathcal{M}) = I \circ \hat{\mathcal{M}} + \mu \circ (1 - \hat{\mathcal{M}})$, where $\hat{\mathcal{M}} \in \mathbb{R}^{w \times h}$ is upsampled \mathcal{M} to a width of w pixels and a height of h pixels.

Existing methods of explaining the output of classifiers [10, 17, 18] or semantic segmentation networks [26] use a fixed value of s for all images, *i.e.*, they fix the size of a *perturbation unit*¹. However, in the case of object detectors, a *perturbation unit* of fixed size can result in perturbations of different sizes to the RoI-pooled features, depending on the size of the proposals, as shown in Figure 1(a). Figure 1(b) shows how the size of a *perturbation unit*, after RoI pooling, can fail to match the sizes of target objects: the perturbations are too coarse for small objects and too fine for large objects. Therefore, we use an adaptive stride $s(a)$ where a is the

¹The *perturbation unit* is a block of image pixels perturbed by a single element of \mathcal{M} .

ratio of the area of the bounding box predicted by the object detector to that of the image, so that we use a small stride for a small object and a large stride for a large object.

3.3. Generating Pseudo Ground Truth

Since the BBAM is a pixel-level localization of the target object in a bounding box predicted by the object detector, it can be used as pseudo ground-truth for weakly supervised semantic and instance segmentation, using the following procedure: We first train an object detector, then create pseudo ground-truth semantic and instance masks for training images, using the BBAM of the trained object detector. These pseudo ground-truth masks can then be used to train semantic and instance segmentation networks. We will now explain this procedure in more detail.

Creating masks. Multiple proposals on a single object yield multiple predictions from the object detector. In order to benefit from the diversity of these predictions, we build the pseudo ground-truth from the BBAMs of multiple proposals. For each ground-truth box, we generate a set of object proposals \mathcal{O} by randomly jittering each coordinate of the box by up to $\pm 30\%$. These proposals are sent to the f^{cls} and the f^{box} . If the f^{cls} correctly predicts the ground-truth class, and the intersection over union (IoU) value associated with the predicted box by f^{box} is greater than 0.8, then the proposal is added to a set of positive proposals $\mathcal{O}^+ \subset \mathcal{O}$. We then use a modified version of $\mathcal{L}_{\text{perturb}}$ in Eq. 1 to amalgamate all the positive proposals into a single localization map, as follows:

$$\begin{aligned} \mathcal{L}_{\text{perturb}} = \mathbb{E}_{o \in \mathcal{O}^+} & [\mathbb{1}_{\text{box}} \|t^c - f^{\text{box}}(\Phi(I, \mathcal{M}), o)\|_1 \\ & + \mathbb{1}_{\text{cls}} \|p^c - f^{\text{cls}}(\Phi(I, \mathcal{M}), o)\|_1]. \end{aligned} \quad (3)$$

In this equation both $\mathbb{1}_{\text{box}}$ and $\mathbb{1}_{\text{cls}}$ are set to 1, since the BBAMs of f^{box} and f^{cls} provide complementary localization results (see Section 5 for details). A BBAM obtained in this way may partially cover the target object because not all pixels of the object are considered by f^{box} and f^{cls} . Therefore we refine the BBAM using CRFs [33], following previous work [2, 31, 60]. Finally, we create pseudo instance-level ground-truth masks by considering the pixels in each BBAM with values greater than a threshold θ to be foreground. We denote such a mask as \mathcal{T} .

The threshold θ controls the size of \mathcal{T} . However, the proportion of pixels in each BBAM which correspond to the foreground will vary, so it may not be appropriate to use a fixed θ . Therefore we introduce two thresholds θ_{fg} and θ_{bg} : pixels whose attribution values are higher than θ_{fg} are considered to be part of the foreground, and pixels whose values are lower than θ_{bg} are considered to be part of the background. The remaining pixels are ignored in the loss computations during training segmentation networks.

Refine with MCG proposals. MCG [49] is an unsupervised mask proposal generator, which is commonly used in

weakly supervised instance segmentation [3, 31, 44, 70, 71]. We can use mask proposals generated by MCG to refine a mask \mathcal{T} . We first select the mask proposal that has the highest IoU with \mathcal{T} . However, that proposal may partially cover the target object. We therefore consider other proposals that are completely contained within \mathcal{T} . More formally, given a set of MCG proposals $\{m_i\}_{i=1}^K$, the refined mask \mathcal{T}_r is derived as follows:

$$\begin{aligned} \mathcal{T}_r &= \bigcup_{i \in S} m_i, \quad \text{where} \\ S &= \{i \mid m_i \subset \mathcal{T}\} \cup \{\underset{i}{\operatorname{argmax}} \operatorname{IoU}(m_i, \mathcal{T})\}. \end{aligned} \quad (4)$$

3.4. Training the Segmentation Network

We now explain the procedure that we use for training the semantic and instance segmentation network.

Instance segmentation. We use Mask R-CNN [23], pre-trained on ImageNet [13]. We use a seed growing technique [2, 28, 36, 37] for pseudo-labeling the pixels ignored during training: Starting with the pixels identified by the initial pseudo ground-truth mask, more of the ignored pixels progressively participate in the loss computation as training proceeds. We refer to Huang *et al.* [28] for more details.

Semantic segmentation. We use DeepLab-v2 [8], pre-trained on the ImageNet [13] dataset. The pseudo labels produced in Section 3.3 can easily be made suitable for semantic segmentation by converting them from instance-level to class-level. Pixels assigned to two or more object classes are ignored during the loss computation.

4. Experiments

4.1. Experimental Setup

Dataset and evaluation metrics. We conducted experiments on the PASCAL VOC [14] and the MS COCO datasets [42]. The PASCAL VOC dataset contains 20 object classes and one background class. Following the same protocol as other recent work on weakly supervised semantic and instance segmentation [1, 3, 27, 60], we used an augmented set of 10,582 training images produced by Hariharan *et al.* [22]. The MS COCO dataset has 118K training images containing 80 object classes. We report mean intersection-over-union (mIoU) values for semantic segmentation. For instance segmentation, we report average precision (AP_τ) at IoU thresholds τ ; averaged AP over IoU thresholds from 0.5 to 0.95; and the average best overlap (ABO).

Reproducibility. We used the PyTorch [48] implementation [45] of Faster R-CNN [52] and Mask R-CNN [23]. For semantic segmentation, we used the PyTorch implementation of DeepLab-v2-ResNet101 [46]. We set $s(a)$ to $16 + 48\sqrt{a}$ and λ to 0.007. We set θ_{fg} and θ_{bg} to 0.8 and 0.2 respectively. To find \mathcal{M}^* in Eq. 1, we used Adam optimizer [32] with a learning rate of 0.02 for 300 iterations. The experiments

Table 1: Weakly supervised instance segmentation performance on PASCAL VOC 2012 *val* images.

Method	AP ₂₅	AP ₅₀	AP ₇₀	AP ₇₅	ABO
Full supervision: Instance masks					
MNC CVPR ’16 [12]	-	63.5	41.5	-	-
Mask R-CNN ICCV ’17 [23]	77.3	69.1	49.9	41.9	65.8
Weak supervision: Image-level tags					
PRM CVPR ’18 [70]	44.3	26.8	-	9.0	37.6
IAM CVPR ’19 [71]	45.9	28.8	-	11.9	41.9
Label-PENet ICCV ’19 [19]	49.1	30.2	-	12.9	41.4
CountSeg CVPR ’19 [9]	48.5	30.2	-	14.4	44.3
IRNet CVPR ’19 [1]	-	46.7	23.5	-	-
Kim <i>et al.</i> WACV ’21 [29]	56.6	38.1	-	12.3	48.2
LIID TPAMI ’20 [44]	-	48.4	-	24.9	50.8
Arun <i>et al.</i> ECCV ’20 [3]	59.1	49.7	29.2	27.1	-
Weak supervision: Bounding boxes					
SDI CVPR ’17 [31]	-	44.8	-	16.3	49.1
Liao <i>et al.</i> ICASSP ’19 [40]	-	51.3	-	22.4	51.9
Sun <i>et al.</i> Access ’20 [62]	-	56.9	-	21.4	56.9
Hsu <i>et al.</i> NeurIPS ’19 [27]	75.0	58.9	30.4	21.6	-
Arun <i>et al.</i> ECCV ’20 [3]	73.1	57.7	33.5	31.2	-
BBAM (Ours)	76.8	63.7	39.5	31.8	63.0

were performed on NVIDIA Tesla V100 GPUs. For MCG mask proposals, we used the pre-computed proposals for PASCAL VOC and MS COCO images provided by Pont-Tuset *et al.* [49].

4.2. Weakly Supervised Instance Segmentation

Results on PASCAL VOC. Table 1 compares the performance of our method with that of other recent methods of weakly supervised instance segmentation which use image-level tags or bounding boxes. Our method significantly outperforms those methods. Specifically, the AP₅₀ and AP₇₀ values of our method are both 6.0% higher than those of the previous best performing method which also uses bounding box annotation [3]. We include results from two fully supervised methods: MNC [12] and Mask R-CNN [23]. The performance of Mask R-CNN [23], which is fully supervised, can be viewed as an upper bound on the achievable performance of our method. We achieve 92.2% and 95.7% of the performance of fully supervised Mask R-CNN, in terms of AP₅₀ and ABO respectively. Figure 2 presents examples of instance masks produced by our method.

Results on MS COCO 2017. This is a challenging dataset containing more objects in an image on average than PASCAL VOC. The sizes of instances of objects are also more diverse. Table 2 compares the performance of our method with that of other weakly supervised instance segmentation methods with various levels of supervision on MS COCO. Our method achieves a 6.7% higher value of AP₇₅ than the previous best performing method which uses bound-

Table 2: Comparison of instance segmentation methods with various types of supervision on MS COCO. The results of Hsu *et al.* [27] were obtained from [here](#).

Method	sup.	AP	AP ₅₀	AP ₇₅
MS COCO <i>val</i> images				
Mask R-CNN ICCV ’17 [23]	\mathcal{F}	35.4	57.3	37.5
Shen <i>et al.</i> CVPR ’19 [57]	\mathcal{I}	6.1	11.7	5.5
Laradji <i>et al.</i> arXiv ’19 [35]	\mathcal{I}, \mathcal{P}	7.8	18.2	8.8
Hsu <i>et al.</i> NeurIPS ’19 [27]	\mathcal{B}	21.1	45.5	17.2
BBAM (Ours)	\mathcal{B}	26.0	50.0	23.9
MS COCO <i>test-dev</i> images				
Mask R-CNN ICCV ’17 [23]	\mathcal{F}	35.7	58.0	37.8
Fan <i>et al.</i> ECCV ’18 [16]	$\mathcal{I}, \mathcal{S}_I$	13.7	25.5	13.5
LIID TPAMI ’20 [44]	\mathcal{I}	16.0	27.1	16.5
BBAM (Ours)	\mathcal{B}	25.7	50.0	23.3

\mathcal{F} —Full, \mathcal{I} —Image label, \mathcal{P} —Point, \mathcal{B} —Box, \mathcal{S}_I —Instance saliency

Table 3: Weakly supervised semantic segmentation on PASCAL VOC 2012 *val* and *test* images.

Method	val	test
Full supervision: Semantic masks		
DeepLab TPAMI ’17 [8]	76.8	76.2
Weak supervision: Image-level tags		
FickleNet CVPR ’19 [36]	64.9	65.3
CIAN AAAI ’20 [15]	64.3	65.3
Chang <i>et al.</i> CVPR ’20 [7]	66.1	65.9
Sun <i>et al.</i> ECCV ’20 [61]	66.2	66.9
Weak Supervision: Bounding boxes		
WSSL ICCV ’15 [47]	60.6	62.2
BoxSup ICCV ’15 [11]	62.0	64.6
SDI CVPR ’17 [31]	69.4	-
Song <i>et al.</i> CVPR ’19 [60]	70.2	-
BBAM (Ours)	73.7	73.7

ing box annotations. Since the labels for *test-dev* images are not publicly available, the results for the *test-dev* images were obtained from the MS COCO challenge website.

4.3. Weakly Supervised Semantic Segmentation

Table 3 compares published mIoU values achieved by recent methods performing semantic segmentation on validation and test images from the PASCAL VOC 2012 dataset. Since the labels for test images are not publicly available, the results for the test images were obtained from the official PASCAL VOC evaluation server. Our method, using the BBAM, yields an mIoU value of 73.7 for both the validation and the test images in the PASCAL VOC 2012 semantic segmentation benchmark. Our method outperforms all the methods that use image-level tags or bounding boxes for supervision. This new state-of-the-art performance was achieved with vanilla DeepLab-v2 [8] without any modifica-



Figure 2: Examples of predicted instance masks for PASCAL VOC *val* images of IRNet [1], Hsu *et al.* [27], and ours.

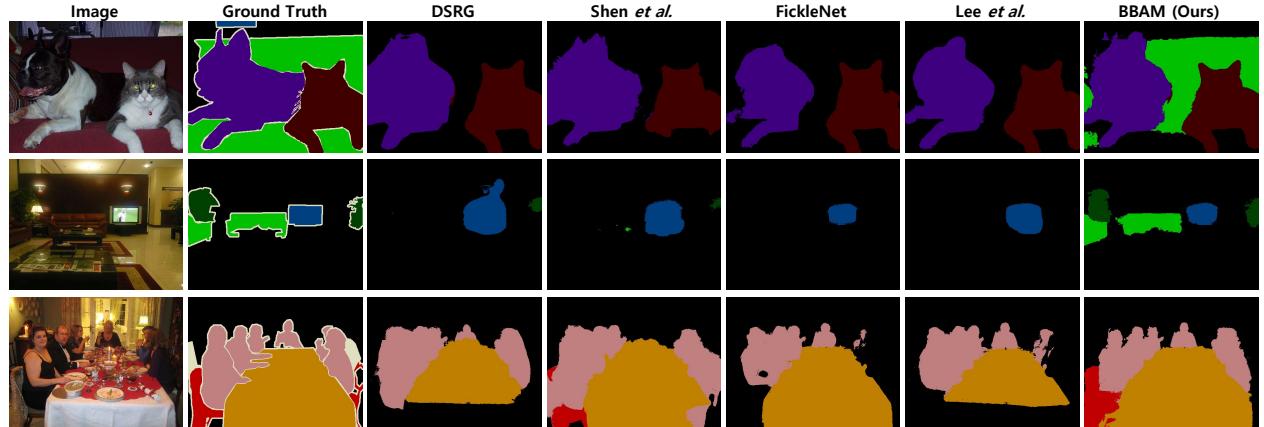


Figure 3: Examples of predicted semantic masks for PASCAL VOC *val* images of DSRG [28], Shen *et al.* [56], FickleNet [36], Lee *et al.* [37], and our method.

Table 4: Effectiveness of using MCG proposals for instance segmentation. AP_S , AP_M , and AP_L respectively denote the AP values for small, medium, and large objects.

MCG	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
PASCAL VOC <i>val</i> images:						
✗	29.6	61.9	25.8	5.6	21.6	40.1
✓	33.4	63.7	31.8	6.5	26.4	44.1
MS COCO <i>val</i> images:						
✗	23.5	47.9	20.3	10.4	24.9	36.5
✓	26.0	50.0	23.9	10.8	28.5	40.3

tions to networks or additional training techniques, such as label refinement during training [11], recursive training [31], or fine-tuning with additional losses [60]. Figure 3 presents examples of semantic masks produced by our method.

The concurrent method, Box2Seg [34], achieved an mIoU of 76.4% on the PASCAL VOC validation images, but it is based on UperNet [67], which is a more powerful segmentation network than DeepLab-v2 [46]. For a fair comparison between Box2Seg [34] and our BBAM, we attempt to relieve the benefit of UperNet [67] over DeepLab-v2 [8] by comparing the relative performance of the weakly supervised model

to the fully supervised model. Box2Seg achieves 88.4% of the performance of its fully supervised equivalent (76.4 vs. 86.4); but the corresponding figure for BBAM and its fully supervised equivalent is 96.7% (73.7 vs. 76.2).

4.4. Ablation Study

MCG proposals. Table 4 shows how mask refinement with MCG proposals improves the instance segmentation performance of our method on the PASCAL VOC and MS COCO datasets. Mask refinement with MCG proposals is particularly effective on masks for medium and large objects. The results obtained without MCG proposals offer the possibility of a fairer comparison with Hsu *et al.* [27], which do not use MCG proposals. Our method produces better results than that of Hsu *et al.* [27] for both the PASCAL VOC and MS COCO datasets, which are shown in Tables 1 and 2 respectively. Hereinafter, to observe the contribution of each component of our system, we report results without using MCG proposals.

Box and cls heads. BBAM can provide a separate attribution map for each head of the object detector by controlling the logical variables 1_{box} and 1_{cls} in Eq. 3. Figure 4 shows the effect of the BBAM obtained from each head on the

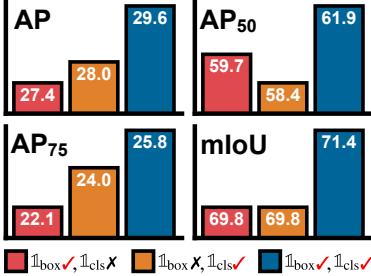


Figure 4: Effect of each head on instance and semantic segmentation.

performance of weakly supervised semantic and instance segmentation. Using the BBAM obtained from either the *box head* ($\mathbb{1}_{\text{box}} = 1$ and $\mathbb{1}_{\text{cls}} = 0$) or the *cls head* ($\mathbb{1}_{\text{box}} = 0$ and $\mathbb{1}_{\text{cls}} = 1$) shows competent performance, but the best performance is achieved when the two heads are used together. We attribute this to the complementary property of the two heads, which is examined in more detail in Section 5.

Parameter sensitivity analysis. Table 5 shows the effect of the thresholds θ_{fg} and θ_{bg} , and the seed growing technique \mathcal{G} . When θ_{fg} equals to θ_{bg} , all pixels are assigned to either the foreground or the background. We see that ignoring some pixels can improve the AP values, and the seed growing technique further improves performance. We then studied the effect of λ , which controls the sparsity of the BBAM, on the performance of weakly supervised semantic and instance segmentation, with the results shown in Table 6. Our method shows similar performance on semantic and instance segmentation over a broad range of values of λ .

5. Detailed Analysis of the BBAM

Examples of BBAMs. Figure 5 shows BBAMs for validation images from PASCAL VOC [14] and MS COCO [42]. The BBAMs have high values on the boundary and discriminative parts of each object, which are informative in conducting object detection.

Complementary operation of the *box* and *cls* heads. To determine which regions of an object are important to each head, we investigated the distribution of high-value pixels in the BBAM produced by each head. In Figure 6(a), \mathcal{C} is the set of points on the contour of the object mask, and \vec{x}_c is its centroid. For each pixel \vec{x} , we determine $r_1 = \|\vec{x} - \vec{x}_c\|_2$ and $r_2 = \min_{\vec{c} \in \mathcal{C}} \|\vec{x} - \vec{c}\|_2$. Letting the angle between $\vec{x} - \vec{x}_c$ and the x -axis be θ , the position of the pixel \vec{x} relative to \vec{x}_c is $\vec{R} = (\frac{r_1}{r_1+r_2} \cos \theta, \frac{r_1}{r_1+r_2} \sin \theta)$. In Figure 6(b), we plot the relative positions of all the pixels with attribution values above 0.9 obtained from validation images of the PASCAL VOC dataset. Pixels for which $\|\vec{R}\|_2 \approx 1$ are near the boundary of the object. We observed that high values attributed by the *box head* mainly occur near the boundary of the object, and those by the *cls head* mainly occur in the interior.

Furthermore, we observed how much the prediction of

θ_{fg}	θ_{bg}	\mathcal{G}	AP	AP ₅₀	AP ₇₅
0.2	0.2	\times	24.8	58.3	18.1
0.5	0.5	\times	28.3	59.5	24.7
0.8	0.8	\times	27.8	59.0	23.3
0.3	0.7	\times	28.1	59.5	24.0
0.3	0.7	\checkmark	28.4	59.6	24.6
0.2	0.8	\times	28.6	60.4	24.0
0.2	0.8	\checkmark	29.6	61.9	25.8

Table 5: Analysis of thresholds θ_{fg} and θ_{bg} , and effect of the growing technique \mathcal{G} .

λ	Ins.		Sem.	
	AP	AP ₅₀	AP ₇₅	mIoU
0.001	26.6	58.7	21.1	67.9
0.003	28.1	59.9	22.8	69.7
0.005	28.7	60.2	24.3	70.8
0.007	29.6	61.9	25.8	71.4
0.010	28.7	60.4	24.4	70.7
0.020	28.3	59.6	23.7	70.3

Table 6: Effect of λ on instance (Ins.) and semantic (Sem.) segmentation.

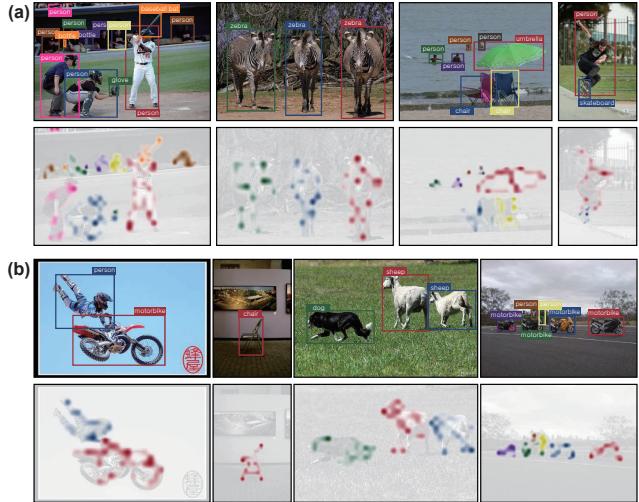


Figure 5: Examples of the predicted boxes and corresponding BBAMs. (a) BBAMs for MS COCO validation images. (b) BBAMs for PACSAL VOC validation images. Each BBAM corresponds to the predicted box of the same color.

each head changes when either of $\mathbb{1}_{\text{box}}$ and $\mathbb{1}_{\text{cls}}$ is set to 1 during the optimization of Eq. 1. The extent of the change in prediction of each head can be inferred from the corresponding loss in Eq. 2. Figure 6(c) shows that applying the optimization of Eq. 1 to one of the heads increases the loss of the other head, implying that the discriminative area of the image necessary for each head is not sufficient for the other head to maintain the prediction. These two observations suggest that the BBAM of each head provides complementary attributions. Examples of BBAMs obtained from each head are presented in the Appendix.

Label noise in object detection. We also looked at the robustness of our system against noisy box coordinate labels in instance segmentation. Hsu *et al.* [27] considered the effect of up to $\pm 15\%$ of label noise: we extend this to $\pm 20\%$. The validity of the bounding box tightness priors used by Hsu *et al.* [27] is seriously compromised by inaccurate box coordinates, with a considerable effect on performance, as shown in Figure 7(a). Our method shows better robustness than that of Hsu *et al.* [27], whether the noise consists of expanded or contracted bounding box annotations.

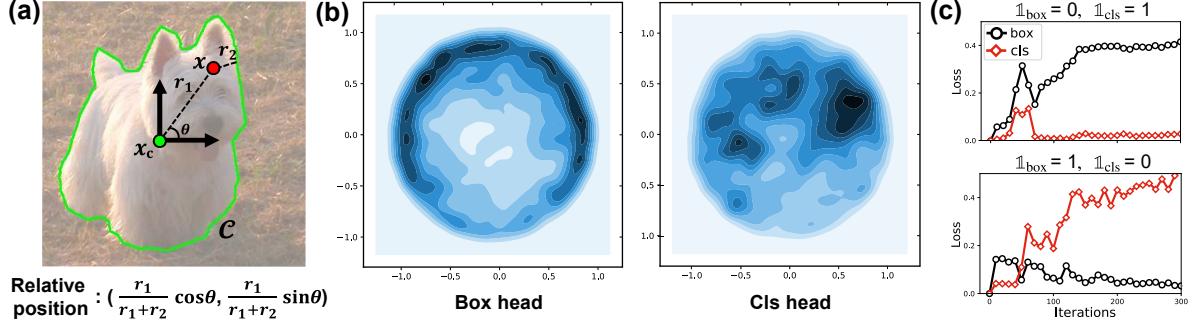


Figure 6: Complementary operation of the *box head* and the *cls head*. (a) The definition of relative position. (b) Relative positions of the highly activated pixels from each head. (c) *Box* and *class* loss curves.

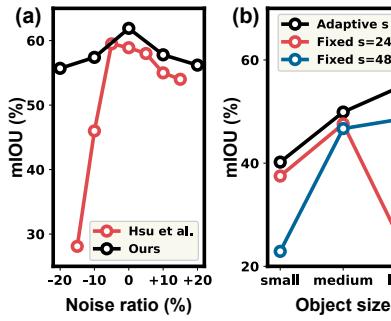


Figure 7: (a) Robustness against noisy box coordinate labels. (b) Localization accuracy by different strides. (c) Localization accuracy by different attribution methods.

Effectiveness of an adaptive stride $s(a)$. As mentioned in Section 3.2, we use an adaptive stride $16 \leq s(a) \leq 64$ to cope with feature transformation due to RoI pooling. Figure 7(b) shows the IoU between the BBAM and ground truth mask on PASCAL VOC validation images, along with the results using fixed strides of 24 and 48. Figure 7(b) shows that a small fixed stride ($s=24$) is ineffective with large objects, as is a large fixed stride ($s=48$) with small objects. By contrast, an adaptive stride $s(a)$ can deal with objects of various sizes.

Comparison with gradient-based methods. Gradient-based attribution methods, such as SimpleGrad [68], SmoothGrad [59], and Grad-CAM [55] can also provide attributions for the output of an object detector. However, since only the subset of features associated with the imperfect proposal is delivered to the *cls* and *box* heads, the gradients with respect to pixels, which exist outside the proposal yet essential for prediction, can vanish (but not completely, due to the receptive field). We provide empirical results supporting this analysis on the PASCAL VOC validation images: (1) Figure 8 shows examples in which SimpleGrad [68] is applied to three similar predictions from different proposals. Pixels outside the proposal do indeed influence the predictions, but SimpleGrad’s attributions mainly appear inside the proposal. (2) We observed that the majority (87%) of pixels with attribution values above 0.9 appear inside the imperfect proposal;

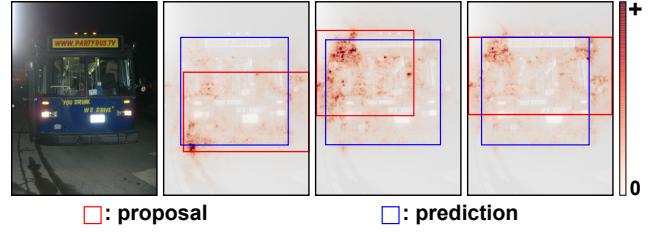


Figure 8: Examples of SimpleGrad [68] for three similar predictions obtained from different proposals.

the mean IoU between the set of positive proposals and the corresponding predictions is low (*i.e.*, 0.56). (3) Figure 7(c) shows that attribution maps from gradient-based attribution methods correlate poorly with ground truth masks.

6. Conclusions

We have introduced a bounding box attribution map (BBAM), which provides pixel-level localization of each target object in its bounding box by finding the smallest region that preserves the predictions of the object detector. Our formulation is built on two-stage object detectors, but applying our method to one-stage object detectors is straightforward as long as they have *box* and *cls* heads. Our experiments demonstrate that the BBAM achieves state-of-the-art performance on the PASCAL VOC and MS COCO benchmarks in weakly supervised semantic and instance segmentation. We have also analyzed BBAMs from various viewpoints, and compared our technique with other attribution methods, to provide a deeper understanding of our approach. We expect BBAMs to be a staple of future work on weakly supervised semantic and instance segmentation with bounding boxes, on a par with the CAM for class labels.

Acknowledgements: This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) [2018R1A2B3001628], AIR Lab (AI Research Lab) in Hyundai & Kia Motor Company through HKMC-SNU AI Consortium Fund, and the Brain Korea 21 Plus Project in 2021.

References

- [1] Jiwoon Ahn, Sunghyun Cho, and Suha Kwak. Weakly supervised learning of instance segmentation with inter-pixel relations. In *CVPR*, 2019.
- [2] Jiwoon Ahn and Suha Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *CVPR*, 2018.
- [3] Aditya Arun, CV Jawahar, and M Pawan Kumar. Weakly supervised instance segmentation by learning annotation consistent instances. In *ECCV*, 2020.
- [4] Amy Bearman, Olga Russakovsky, Vittorio Ferrari, and Li Fei-Fei. What’s the point: Semantic segmentation with point supervision. In *ECCV*, 2016.
- [5] Miriam Bellver, Amaia Salvador, Jordi Torres, and Xavier Giro-i Nieto. Budget-aware semi-supervised semantic and instance segmentation. *CVPR Workshops*, 2019.
- [6] Chun-Hao Chang, Elliot Creager, Anna Goldenberg, and David Duvenaud. Explaining image classifiers by counterfactual generation. *ICLR*, 2019.
- [7] 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 *CVPR*, 2020.
- [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 TPAMI*, 2017.
- [9] Hisham Cholakkal, Guolei Sun, Fahad Shahbaz Khan, and Ling Shao. Object counting and instance segmentation with image-level supervision. In *CVPR*, 2019.
- [10] Piotr Dabkowski and Yarin Gal. Real time image saliency for black box classifiers. In *NeurIPS*, 2017.
- [11] Jifeng Dai, Kaiming He, and Jian Sun. Boxesup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *ICCV*, 2015.
- [12] Jifeng Dai, Kaiming He, and Jian Sun. Instance-aware semantic segmentation via multi-task network cascades. In *CVPR*, 2016.
- [13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.
- [14] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 2010.
- [15] Junsong Fan, Zhaoxiang Zhang, and Tieniu Tan. Cian: Cross-image affinity net for weakly supervised semantic segmentation. *AAAI*, 2020.
- [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 *ECCV*, 2018.
- [17] Ruth Fong, Mandela Patrick, and Andrea Vedaldi. Understanding deep networks via extremal perturbations and smooth masks. In *ICCV*, 2019.
- [18] Ruth C Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *ICCV*, 2017.
- [19] Weifeng Ge, Sheng Guo, Weilin Huang, and Matthew R Scott. Label-penet: Sequential label propagation and enhancement networks for weakly supervised instance segmentation. In *ICCV*, 2019.
- [20] Ross Girshick. Fast r-cnn. In *ICCV*, 2015.
- [21] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *ICLR*, 2014.
- [22] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *ICCV*, 2011.
- [23] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017.
- [24] Seunghoon Hong, Donghun Yeo, Suha Kwak, Honglak Lee, and Bohyung Han. Weakly supervised semantic segmentation using web-crawled videos. In *CVPR*, 2017.
- [25] Qibin Hou, PengTao Jiang, Yunchao Wei, and Ming-Ming Cheng. Self-erasing network for integral object attention. In *NeurIPS*, 2018.
- [26] Lukas Hoyer, Mauricio Munoz, Prateek Katiyar, Anna Khoreva, and Volker Fischer. Grid saliency for context explanations of semantic segmentation. In *NeurIPS*, 2019.
- [27] Cheng-Chun Hsu, Kuang-Jui Hsu, Chung-Chi Tsai, Yen-Yu Lin, and Yung-Yu Chuang. Weakly supervised instance segmentation using the bounding box tightness prior. In *NeurIPS*, 2019.
- [28] Zilong Huang, Xinggang Wang, Jiasi Wang, Wenyu Liu, and Jingdong Wang. Weakly-supervised semantic segmentation network with deep seeded region growing. In *CVPR*, 2018.
- [29] Jaedong Hwang, Seohyun Kim, Jeany Son, and Bohyung Han. Weakly supervised instance segmentation by deep community learning. In *WACV*, 2021.
- [30] Peng-Tao Jiang, Qibin Hou, Yang Cao, Ming-Ming Cheng, Yunchao Wei, and Hongkai Xiong. Integral object mining via online attention accumulation. In *ICCV*, 2019.
- [31] Anna Khoreva, Rodrigo Benenson, Jan Hosang, Matthias Hein, and Bernt Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In *CVPR*, 2017.
- [32] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ICLR*, 2015.
- [33] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *NeurIPS*, 2011.
- [34] Viveka Kulharia, Siddhartha Chandra, Amit Agrawal, Philip Torr, and Ambrish Tyagi. Box2seg: Attention weighted loss and discriminative feature learning for weakly supervised segmentation. In *ECCV*, 2020.
- [35] Issam H Laradji, Negar Rostamzadeh, Pedro O Pinheiro, David Vázquez, and Mark Schmidt. Instance segmentation with point supervision. *arXiv preprint arXiv:1906.06392*, 2019.
- [36] Jungbeom Lee, Eunji Kim, Sungmin Lee, Jangho Lee, and Sungroh Yoon. Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference. In *CVPR*, 2019.
- [37] Jungbeom Lee, Eunji Kim, Sungmin Lee, Jangho Lee, and Sungroh Yoon. Frame-to-frame aggregation of active regions in web videos for weakly supervised semantic segmentation. In *ICCV*, 2019.
- [38] Sungmin Lee, Jangho Lee, Jungbeom Lee, Chul-Kee Park, and Sungroh Yoon. Robust tumor localization with pyramid grad-cam. *arXiv preprint arXiv:1805.11393*, 2018.
- [39] Kunpeng Li, Ziyan Wu, Kuan-Chuan Peng, Jan Ernst, and Yun Fu. Guided attention inference network. *IEEE TPAMI*,

- 2019.
- [40] Shisha Liao, Yongqing Sun, Chenqiang Gao, Pranav Shenoy KP, Song Mu, Jun Shimamura, and Atsushi Sagata. Weakly supervised instance segmentation using hybrid networks. In *ICASSP*, 2019.
- [41] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017.
- [42] 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*, 2014.
- [43] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *ECCV*, 2016.
- [44] 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. *TPAMI*, 2020.
- [45] Francisco Massa and Ross Girshick. maskrcnn-benchmark: Fast, modular reference implementation of Instance Segmentation and Object Detection algorithms in PyTorch. <https://github.com/facebookresearch/maskrcnn-benchmark>, 2018.
- [46] Kazuto Nakashima. DeepLab with PyTorch. <https://github.com/kazuto1011/deeplab-pytorch>.
- [47] 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*, 2015.
- [48] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.
- [49] Jordi Pont-Tuset, Pablo Arbelaez, Jonathan T Barron, Ferran Marques, and Jitendra Malik. Multiscale combinatorial grouping for image segmentation and object proposal generation. *IEEE TPAMI*, 2016.
- [50] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, 2016.
- [51] Emily Reif, Ann Yuan, Martin Wattenberg, Fernanda B Viégas, Andy Coenen, Adam Pearce, and Been Kim. Visualizing and measuring the geometry of bert. In *NeurIPS*, 2019.
- [52] Shaoting Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015.
- [53] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. " grabcut" interactive foreground extraction using iterated graph cuts. *ACM transactions on graphics (TOG)*, 2004.
- [54] Karl Schulz, Leon Sixt, Federico Tombari, and Tim Landgraf. Restricting the flow: Information bottlenecks for attribution. *ICLR*, 2020.
- [55] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 2017.
- [56] Tong Shen, Guosheng Lin, Chunhua Shen, and Ian Reid. Bootstrapping the performance of webly supervised semantic segmentation. In *CVPR*, 2018.
- [57] Yunhang Shen, Rongrong Ji, Yan Wang, Yongjian Wu, and Liujuan Cao. Cyclic guidance for weakly supervised joint detection and segmentation. In *CVPR*, 2019.
- [58] Wataru Shimoda and Keiji Yanai. Self-supervised difference detection for weakly-supervised semantic segmentation. In *ICCV*, 2019.
- [59] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.
- [60] 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 *CVPR*, 2019.
- [61] Guolei Sun, Wenguan Wang, Jifeng Dai, and Luc Van Gool. Mining cross-image semantics for weakly supervised semantic segmentation. In *ECCV*, 2020.
- [62] Yongqing Sun, Shisha Liao, Chenqiang Gao, Chengjuan Xie, Feng Yang, Yue Zhao, and Atsushi Sagata. Weakly supervised instance segmentation based on two-stage transfer learning. *IEEE Access*, 2020.
- [63] Meng Tang, Abdelaziz Djelouah, Federico Perazzi, Yuri Boykov, and Christopher Schroers. Normalized cut loss for weakly-supervised cnn segmentation. In *CVPR*, 2018.
- [64] 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 *CVPR*, pages 1568–1576, 2017.
- [65] Yunchao Wei, Huixin 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 *CVPR*, 2018.
- [66] Tianfu Wu and Xi Song. Towards interpretable object detection by unfolding latent structures. In *ICCV*, 2019.
- [67] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 418–434, 2018.
- [68] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *ECCV*, 2014.
- [69] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016.
- [70] Yanzhao Zhou, Yi Zhu, Qixiang Ye, Qiang Qiu, and Jianbin Jiao. Weakly supervised instance segmentation using class peak response. In *CVPR*, 2018.
- [71] Yi Zhu, Yanzhao Zhou, Huijuan Xu, Qixiang Ye, David Doermann, and Jianbin Jiao. Learning instance activation maps for weakly supervised instance segmentation. In *CVPR*, 2019.

Table A1: Comparison of per-class mIoU scores.

	bkg	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	mIOU
Results on validation images:																						
Shen <i>et al.</i> [56]	86.8	71.2	32.4	77.0	24.4	69.8	85.3	71.9	86.5	27.6	78.9	40.7	78.5	79.1	72.7	73.1	49.6	74.8	36.1	48.1	59.2	63.0
CIAN [15]	88.2	79.5	32.6	75.7	56.8	72.1	85.3	72.9	81.7	27.6	73.3	39.8	76.4	77.0	74.9	66.8	46.6	81.0	29.1	60.4	53.3	64.3
FickleNet [36]	89.5	76.6	32.6	74.6	51.5	71.1	83.4	74.4	83.6	24.1	73.4	47.4	78.2	74.0	68.8	73.2	47.8	79.9	37.0	57.3	64.6	64.9
SSDD [58]	89.0	62.5	28.9	83.7	52.9	59.5	77.6	73.7	87.0	34.0	83.7	47.6	84.1	77.0	73.9	69.6	29.8	84.0	43.2	68.0	53.4	64.9
Lee <i>et al.</i> [37]	90.8	82.2	35.1	82.4	72.2	71.4	82.7	75.0	86.9	18.3	74.2	29.6	81.1	79.2	74.7	76.4	44.2	78.6	35.4	72.8	63.0	66.5
BBAM (Ours)	92.7	80.6	33.8	83.7	64.9	75.5	91.3	80.4	88.3	37.0	83.3	62.5	84.6	80.8	74.7	80.0	61.6	84.5	48.6	85.8	71.8	73.7
Results on test images:																						
Shen <i>et al.</i> [56]	87.2	76.8	31.6	72.9	19.1	64.9	86.7	75.4	86.8	30.0	76.6	48.5	80.5	79.9	79.7	72.6	50.1	83.5	48.3	39.6	52.2	63.9
FickleNet [36]	90.3	77.0	35.2	76.0	54.2	64.3	76.6	76.1	80.2	25.7	68.6	50.2	74.6	71.8	78.3	69.5	53.8	76.5	41.8	70.0	54.2	65.0
SSDD [58]	89.0	62.5	28.9	83.7	52.9	59.5	77.6	73.7	87.0	34.0	83.7	47.6	84.1	77.0	73.9	69.6	29.8	84.0	43.2	68.0	53.4	64.9
Lee <i>et al.</i> [37]	91.2	84.2	37.9	81.6	53.8	70.6	79.2	75.6	82.3	29.3	76.2	35.6	81.4	80.5	79.9	76.8	44.7	83.0	36.1	74.1	60.3	67.4
BBAM (Ours)	92.8	83.5	33.4	88.9	61.8	72.8	90.3	83.5	87.6	34.7	82.9	66.1	83.9	81.1	78.3	77.4	55.2	86.7	58.5	81.5	66.4	73.7

A. Appendix

A.1. Implementation details

TV norm. To suppress the artifacts in the mask \mathcal{M} , we regularized \mathcal{M} with total variation (TV) norm in Eq. 1 in the main paper, as done in Fong *et al.* [18]. The resulting loss function to find the best \mathcal{M}^* becomes:

$$\begin{aligned}\mathcal{L}_{\mathcal{M}} = & \lambda \|\mathcal{M}\|_1 + \lambda_{\text{TV}} \|\nabla \mathcal{M}\|_{\beta} \\ & + \mathbb{1}_{\text{box}} \|t^c - f^{\text{box}}(\Phi(I, \mathcal{M}), o)\|_1 \\ & + \mathbb{1}_{\text{cls}} \|p^c - f^{\text{cls}}(\Phi(I, \mathcal{M}), o)\|_1,\end{aligned}\quad (5)$$

where λ_{TV} is a balancing factor for TV norm. We set λ_{TV} to 10^{-4} and β to 3. We observed that the resulting mask \mathcal{M}^* has a little dependency on the value of λ_{TV} .

We can find the best \mathcal{M}^* by using gradient descent with respect to \mathcal{M} . Letting the mask at iteration t be \mathcal{M}^t , the mask at iteration $t+1$ can be expressed as

$$\mathcal{M}^{t+1} = \mathcal{M}^t - \xi \nabla_{\mathcal{M}^t} \mathcal{L}_{\mathcal{M}^t}, \quad (6)$$

where ξ is a learning rate. Indeed, the update in Eq 6 was implemented through Adam optimizer.

Optimization details for semantic segmentation. We used the default setting provided by [46], except for the batch size, the number of training iterations, and the learning rate. We set the batch size to 8, the number of training iterations to 2.4×10^4 , and the learning rate to 2×10^{-4} .

Optimization details for instance segmentation on the PASCAL VOC dataset. Regarding the characteristics of the PASCAL VOC dataset [14], we adjusted the input image size and the anchor size accordingly. We set the max and min size of training images to 800 and 512, respectively, and anchor sizes for each FPN level to [21, 42, 84, 168, 332]. We trained Mask R-CNN [23] with a learning rate 8×10^{-3} for 2×10^4 iterations.

Optimization details for instance segmentation on the MS COCO 2017 dataset. We followed the default settings provided by maskrcnn-benchmark repository [45].

Post-processing of semantic and instance segmentation. CRF [33] is a popular post-processing technique for semantic and instance segmentation [27, 28, 36, 37, 58]. We also used CRFs as a post-processing method for semantic and instance segmentation.

A.2. Additional Results

Comparison of per-class mIoU scores. Table A1 shows the per-class mIoU of our method and recently produced methods.

More examples of BBAMs. We present more examples of BBAMs for PASCAL VOC [14] validation images with Faster R-CNN [52] (Figure A1) and for MS COCO 2017 [42] validation images with Faster R-CNN [52] (Figure A2).

Additional mask examples on semantic segmentation. Figure A3 shows more examples of the semantic masks produced by DSRG [28], Shen *et al.* [56], FickleNet [36], Lee *et al.* [37], and our method.

More mask examples on instance segmentation. Figure A4 shows more examples of the instance masks on PASCAL VOC 2012 validation images obtained from IRNet [1], Hsu *et al.* [27], and our method. Figure A5 shows examples of instance masks on MS COCO 2017 validation images obtained by our method.

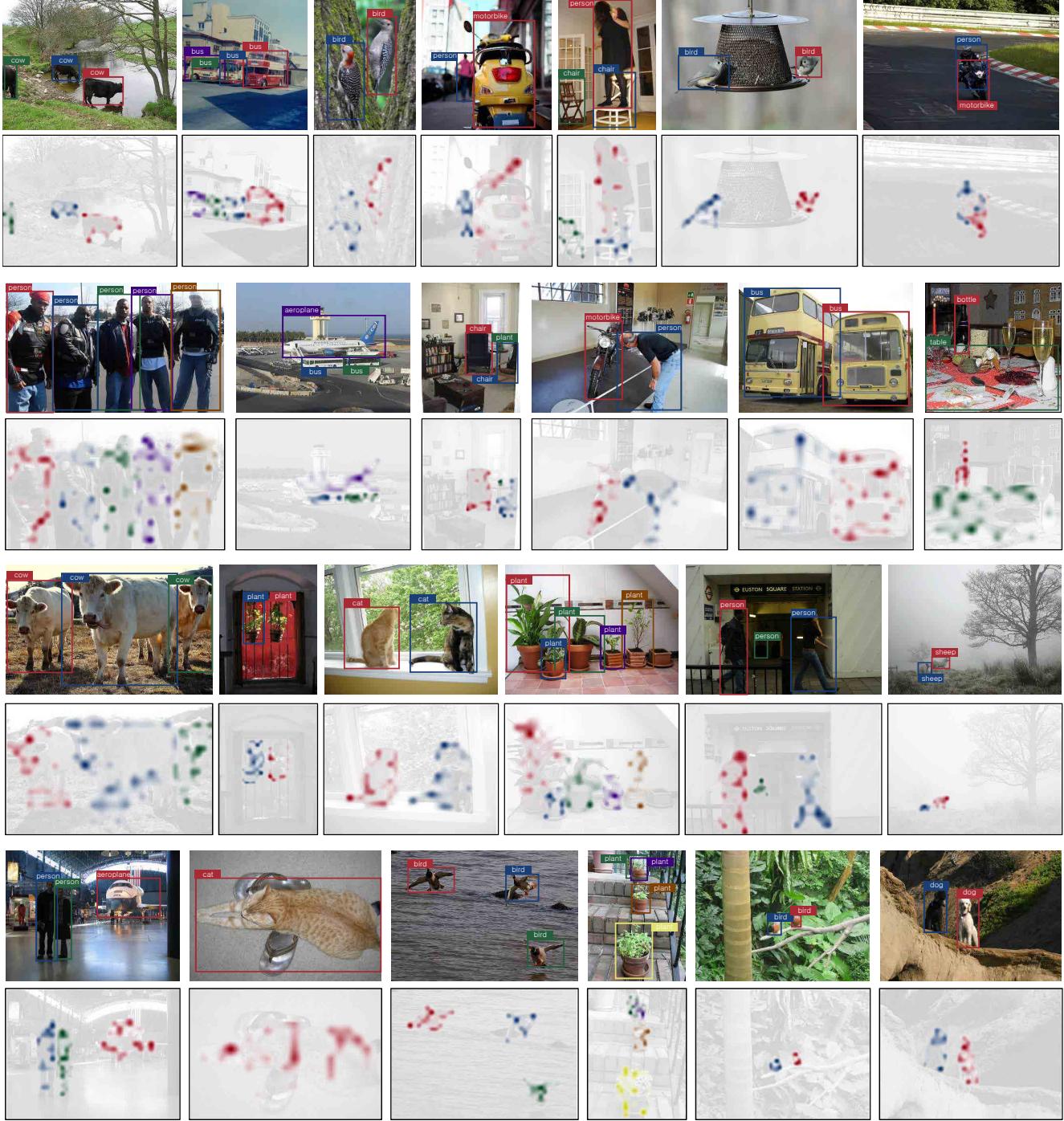


Figure A1: Examples of PASCAL VOC [14] validation images with the results of object detection and corresponding BBAMs, obtained from Faster R-CNN [52].

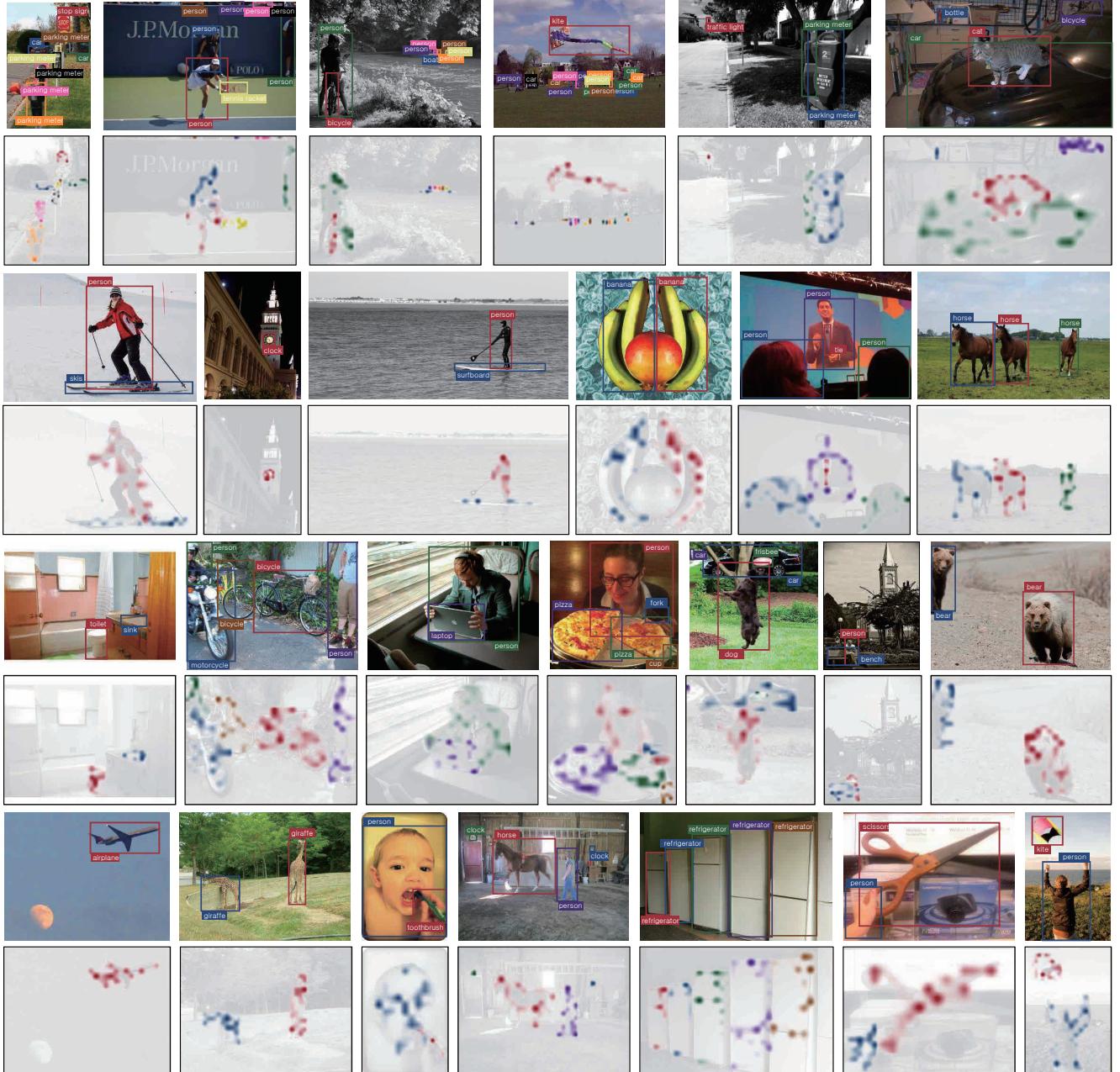


Figure A2: Examples of MS COCO 2017 [42] validation images with the results of object detection and corresponding BBAMs, obtained from Faster R-CNN [52].

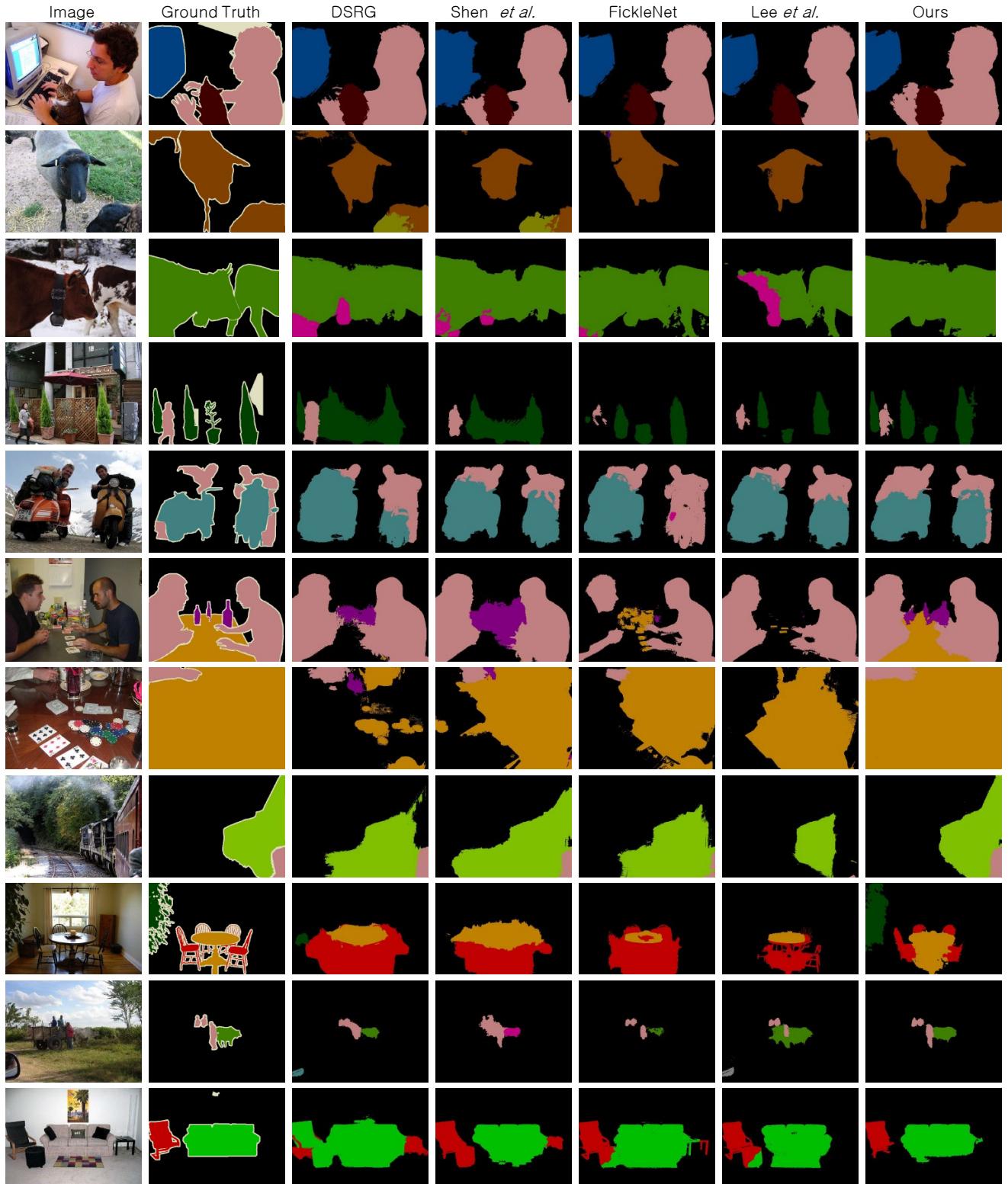


Figure A3: Examples of predicted semantic masks for PASCAL VOC validation images of DSRG [28], Shen *et al.* [56], FickleNet [36], Lee *et al.* [37], and our method.

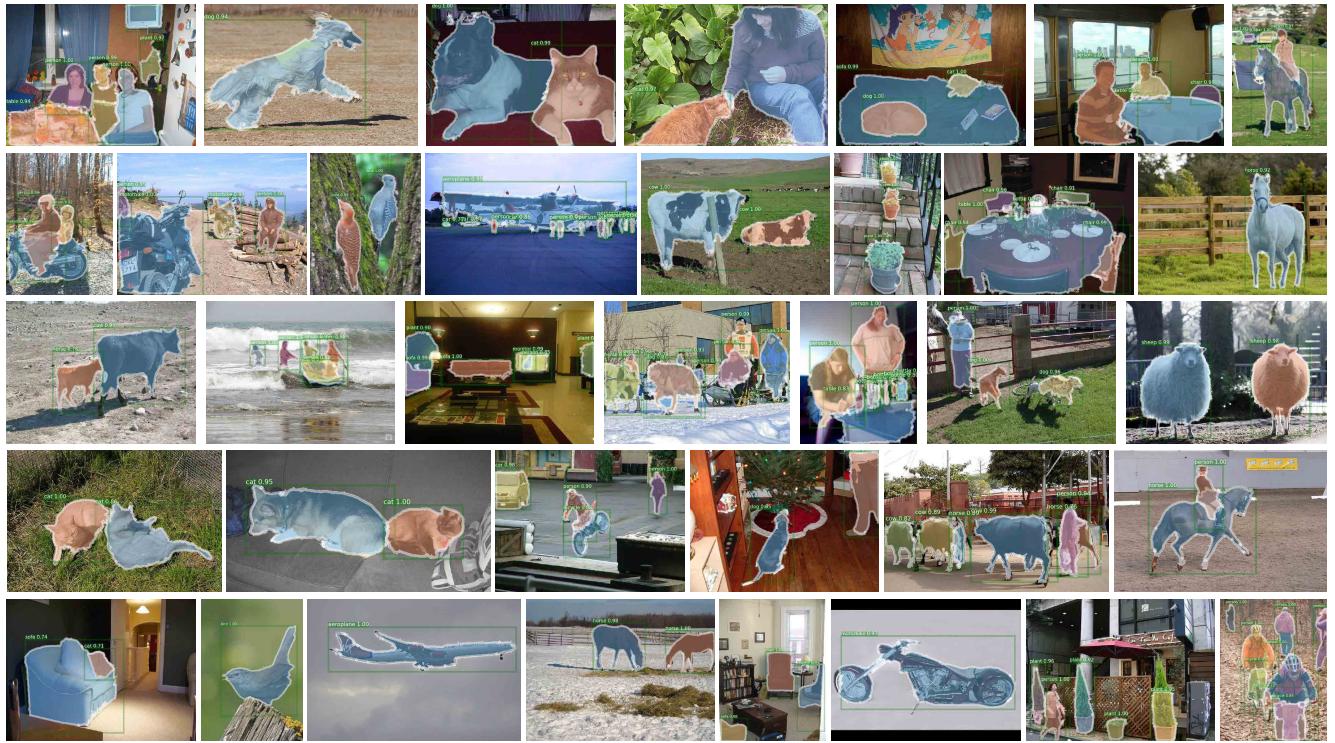


Figure A4: Examples of predicted instance masks for PASCAL VOC validation images of our method.

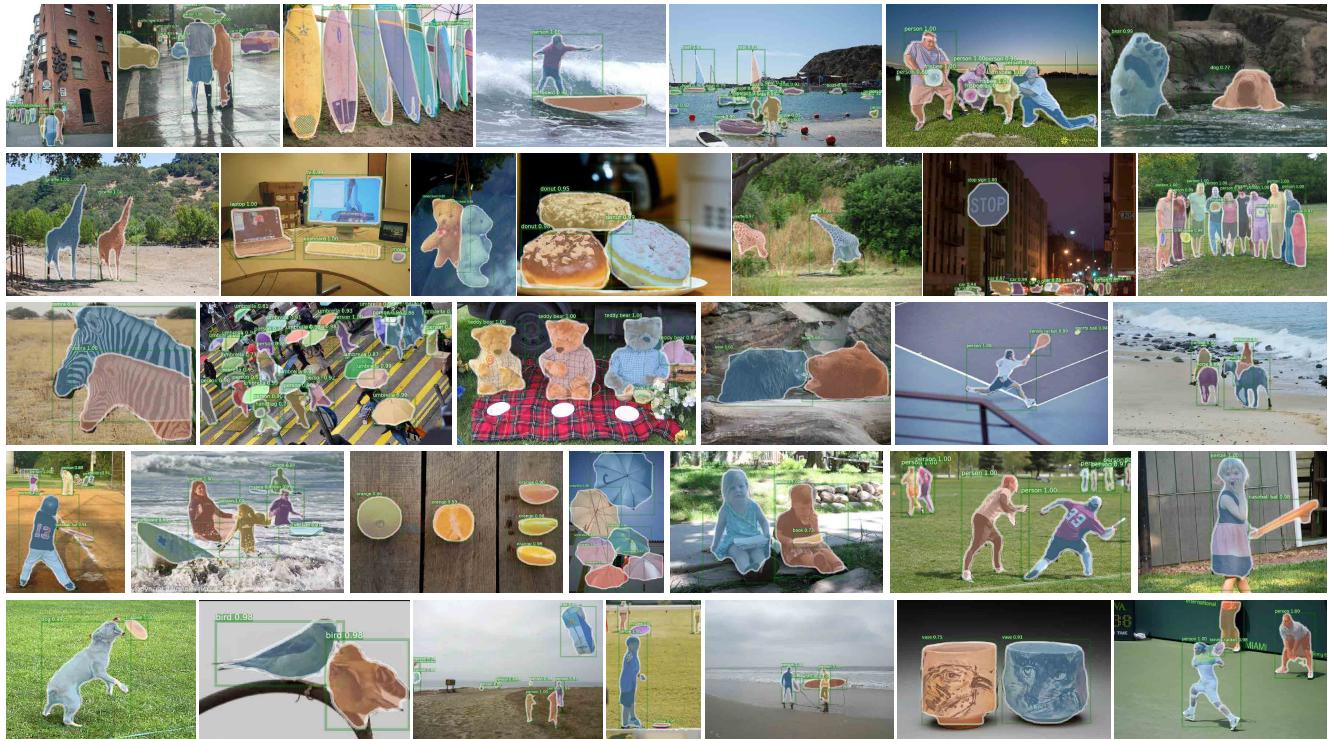


Figure A5: Examples of predicted instance masks for MS COCO 2017 validation images of our method.