Where are the Masks: Instance Segmentation with Image-level Supervision

Issam H. Laradji[1,2]
issamou@cs.ubc.ca
David Vazquez[1]
dvazquez@elementai.com
Mark Schmidt[2]
schmidtm@cs.ubc.ca

[1] Element AI
Montreal, Canada
[2] University of British Columbia
Vancouver, Canada

Abstract

A major obstacle in instance segmentation is that existing methods often need many per-pixel labels in order to be effective. These labels require large human effort and for certain applications, such labels are not readily available. To address this limitation, we propose a novel framework that can effectively train with image-level labels, which are significantly cheaper to acquire. For instance, one can do an internet search for the term "car" and obtain many images where a car is present with minimal effort. Our framework consists of two stages: (1) train a classifier to generate pseudo masks for the objects of interest; (2) train a fully supervised Mask R-CNN on these pseudo masks. Our two main contribution are proposing a pipeline that is simple to implement and is amenable to different segmentation methods; and achieves new state-of-the-art results for this problem setup. Our results are based on evaluating our method on PASCAL VOC 2012, a standard dataset for weakly supervised methods, where we demonstrate major performance gains compared to existing methods with respect to mean average precision.

1 Introduction

The recent progress in Deep Neural Networks (DNNs) and segmentation frameworks has given us major improvements in the task of instance segmentation [7, 16]. Their success was demonstrated in various applications such as autonomous driving [10], scene understanding [13, 25], and medical imaging [20, 33]. Nonetheless, these methods require a large number of training data with per-pixel labels, or labels which distinguish between object categories and instances in the image. As acquiring them is often prohibitively expensive, the effectiveness of these methods is limited to a small range of datasets and object categories.

Many weakly supervised methods emerged to overcome the need for per-pixel labels. Instead, they only require weaker labels ranging from bounding boxes [18], scribbles [24], and image-level [2, 9, 44] annotations. This makes acquiring datasets a significantly more scalable endeavour. According to Bearman et al. [4], it requires 20 sec/img to acquire image-level labels (which are labels that only indicate whether an object class appears in an image) for PASCAL VOC [13], compared to 239.7 sec/img for acquiring per-pixel labels. To date,

only two weakly supervised methods address instance segmentation with image-level labels, making our work one of the few that tackles a relatively unexplored research area.

Perhaps the first work to address this problem setup is PRM [44]. It trains a classifier which can then identify local regions belonging to different objects of the same category. It extends CAM-based methods [1, 39] by not only identifying large regions where objects are vaguely located, but also identifying peaks that represent the specific locations of the object instances. At test time, the trained PRM obtains the object masks in two steps. First, it uses the gradient with respect to the input to get a rough mask of the objects using a process called peak backpropagation. This results in a peak response map. Then, the masks in this map are replaced by the best matching proposal masks, which are generated from a pretrained object proposal method [3, 32]. However, their results are much worse than that of fully supervised methods, leaving a large room for improvement (Table 1).

Our Weakly-supervised Instance SEgmentation method (WISE) builds on PRM by using its output pseudo masks to train a fully-supervised method, namely, Mask R-CNN [29]. Our intuition as to why this procedure is effective is that Mask R-CNN is potentially robust to noisy pseudo masks, and the noisy labels within these masks might be ignored during training as they are potentially uncorrelated. The success of such a de-noising strategy has been demonstrated in semantic segmentation and object localization [18].

We show that simple techniques for obtaining the pseudo masks lead to a surprisingly effective supervision for Mask R-CNN. We summarize our contributions as follows. (1) We present a novel framework that can effectively train a fully supervised method on pseudo mask labels obtained from image-level class labels; (2) we show that our framework is amenable to different localization and segmentation methods (for example, a density-based PRM [4] can be used for localization and RetinaMask [14] can be used for instance segmentation), and (3) we achieve new state-of-the-art results on a standard weakly-supervised instance segmentation benchmark.

2 Related Work

Instance segmentation is widely studied within the computer vision community [7, 14, 16]. However, an ongoing challenge is that it is time-consuming and expensive to obtain the required per-pixel labels needed by most instance segmentation methods [10, 13]. Current trends to overcome this issue leverage weaker labels (such as image-level labels) and pseudo labels obtained with the help of object proposal methods. While most of these methods are for object detection and semantic segmentation, we review them below as they are relevant.

Instance segmentation. Instance segmentation is one the most challenging tasks in computer vision. The task is to classify every object pixel into its corresponding category and distinguish between object instances [36, 38]. Most recent methods rely on deep networks and follow a two step procedure [7, 14, 16], where they first detect objects and then segment them. For instance, Mask-RCNN [16] uses Faster-RCNN [37] for detection and an FCN network [27] for segmentation. In this work, we use Mask R-CNN as our fully supervised method and train it on pseudo masks instead of the costly per-pixel labels.

Learning with weak supervision. Due to the taxing task of acquiring per-pixel labels, many weakly supervised methods emerged that can leverage labels that are much cheaper to acquire [10, 13]. These labels range from bounding boxes [18], scribbles [24], points [4, 22],

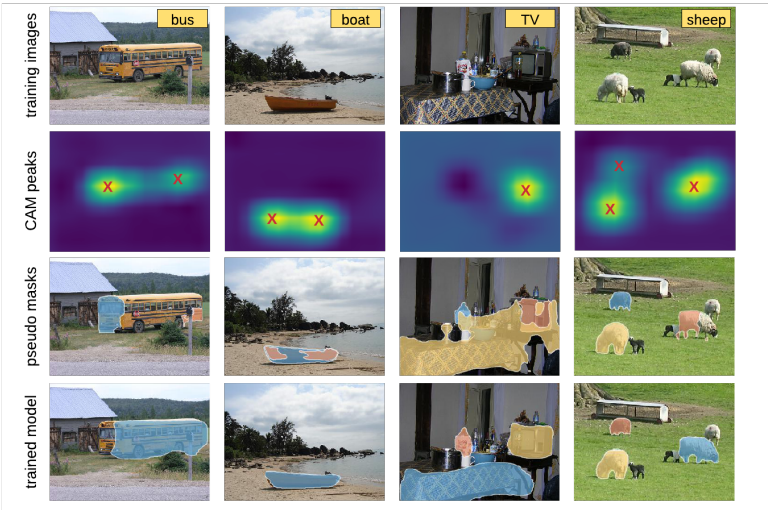

Figure 1: **Framework overview.** Our Weakly-supervised Instance SEgmentation (WISE) method learns to perform instance segmentation with image-level supervision. First, a classifier is trained with a peak stimulation layer to identify peaks at which the objects are located (row 2). A proposal gallery (such as MCG [3]) is used to obtain rough masks for the located objects, which are then used as pseudo masks to train Mask R-CNN [16] (row 3). Row 4 shows the output of a Mask R-CNN trained on the noisy pseudo mask labels.

23], and image-level annotation [44]. Our setup considers one of the weakest forms of annotation, image-level labels.

Image-level labels as weak supervision. Acquiring image-level labels is an attractive form of annotation due to its extremely cheap cost. The annotator only needs to indicate whether a certain object class appears in an image, regardless of how many of them appear. While this form of annotation has gained steam within the research community, most of the proposed methods are for semantic segmentation [1, 15, 41]. Perhaps the lack of such research for instance segmentation is partially accounted for by the fact that instance segmentation is a more challenging task. Only recently did few works emerge for this problem setup [2, 9, 44]. Zhou et al. [44] and Cholakkal et al. [9] extend the Class Activation Map (CAM) [43], by not only identifying a heatmap that vaguely represents the regions where objects are located, but also identifying peaks of that heatmap that represent the locations of different objects. At test time, they adopt a post-processing step that matches each located object with a proposal, generated from an object proposal method. These proposals are considered as the final instance segmentation output. In contrast, we use these outputs as pseudo masks to train a fully supervised method.

Learning with pseudo labels. Our method adopts the following pipeline, generate pseudo-labels and then training a model on these labels in a fully-supervised manner. While this is novel for instance segmentation, similar approaches were used for object detection [40] and semantic segmentation [11, 18, 35] in weakly supervised settings. However, these methods cannot be directly applied to instance segmentation, as they do not distinguish between object instances. Many such methods also rely on object proposals [17] to ease the task of detec-

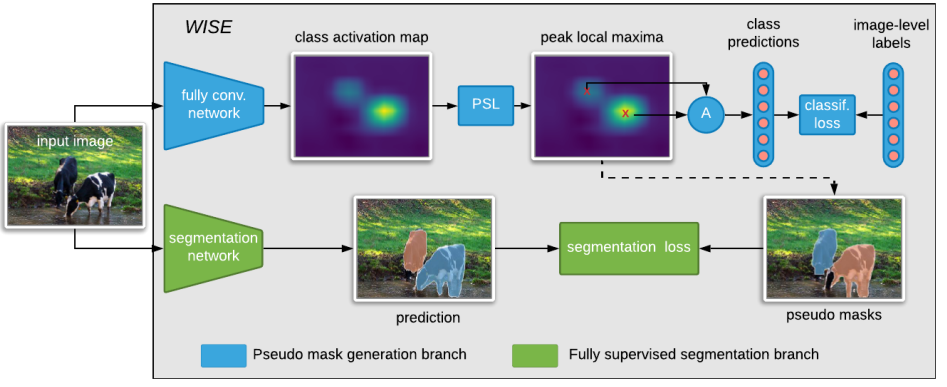

Figure 2: **WISE training.** The first component (shown in blue) learns to classify the images in the dataset. The classifier first outputs a class activation map (CAM); then, obtains CAM's local maximas using a peak stimulation layer (PSL). To train the classifier, the classification loss is computed using the average of these local maximas. As the CAM peaks represent located objects, we select a proposal for each of these objects to obtain pseudo masks. The second component (shown in green) trains a Mask R-CNN on these pseudo masks.

tion [5, 41], and segmentation [9, 19, 30, 44]. Object proposals are class-agnostic methods that can output thousands of object candidates per image and have progressed significantly over the last decade [3, 28, 31, 32, 42, 45]. Similar to PRM [44] and PRM+Density [9], we also leverage object proposals to generate the pseudo masks.

3 Proposed Method

Our approach to instance segmentation with image-level supervision consists of two main steps: (1) obtain pseudo masks for the training images given their ground-truth image-level labels; and (2) train a fully supervised instance segmentation method on these pseudo masks (Figure 2). In particular, this framework is based on two components: a network that obtains the pseudo masks by training a PRM [44] on the image-level labels and leveraging object proposal methods [3]; and a Mask R-CNN [16] as a fully supervised instance segmentation method. We show the training steps of our framework in Algorithm 1.

At test time, we can predict the object instance masks using the trained Mask R-CNN only, discarding the PRM component. In this setup, we are interested in segmenting C classes of objects. For a training image, the image-level label is given as $Y = [y_1, y_2, ..., y_C]$, where $y_i = 1$ or 0, indicating whether the image has an object of class i. We describe our components in more detail below, and also investigate a post-processing steps that can improve Mask R-CNN's final output.

3.1 Pseudo Mask Generation Branch

We rely on PRM [44] to generate segmentation seeds that identify salient parts of the objects. These seeds help in generating pseudo masks as a source of supervision for Mask R-CNN. Following PRM's methodology, we train a CAM-based classifier which has a fully

Algorithm 1 WISE training

Train a CAM-based classifier C until convergence as in PRM [44];
while $iter < max\_iter$ **do**
  Randomly sample a training image I;
  Generate a set of proposals P for I;
  Use PSL on C to obtain the set of peaks L for I;
  Initialize an empty list of Targets T;
  for $(i_k, j_k) \in L$ **do**
    Select a proposal (G_k, b_k) randomly using Eq. 1, it has to intersect with (i_k, j_k);
    Add G_k to list T;
  end for
  Compute $\mathcal{L}(I, T, \theta)$ as in Eq. 2;
  Update the weights for θ using back-propagation;
end while

convolutional network (FCN) followed by a peak stimulation layer (PSL). As shown in Figure 2, the FCN outputs a class activation map (CAM) which specifies the class confidence at each image location. Then, PSL outputs N^c local maximas of CAM within a window size r, namely, $L^c = \{(i_1, j_1), (i_2, j_2), ..., (i_{N^c}, j_{N^c})\}$ which represents locations in the CAM for the c-th object class (more details in Zhou et al. [44]). In order to boost the activations of these local maximas, their average activation is first computed as, $s^c = \frac{1}{N^c} \sum_{(i_k, j_k) \in L^c} M^c_{i_k, j_k}$, where M^c is the activation map corresponding to class c. This average is then used for binary classification, specifically the multi-label soft-margin loss [21]. This classification component is trained until convergence.

We then generate the pseudo masks for the training images by using the trained classifier and an off-the-shelf object proposal method (specified as the dotted line in Figure 2). The peaks obtained from PSL, which represent object locations in the image, are replaced with proposal masks based on their "objectness", which are scores given by the proposal method as confidence measure for being objects. We adopt a de-noising strategy where we select a proposal randomly based on its objectness score: proposals with higher objectness are more likely to be selected. More formally, to obtain the mask for an object located at peak (i, j), we first generate a set of n proposals whose masks intersect with (i, j), namely, $P = \{(G_1, b_1), (G_2, b_2), ..., (G_n, b_n)\}$ with mask G_k and objectness score b_k. Then, the probability of selecting a proposal mask G_k is,

$$P(G_k) = \frac{b_k}{\sum_{j=1}^{n} b_j} \tag{1}$$

The rationale behind selecting proposals randomly is that they have common pixels that correspond to the salient parts of the located object, despite the fact that they have different objectness. While proposal masks are not originally associated with a class label, we get the object class label information from CAM and assign it to the corresponding proposals. These proposals can be used as pseudo masks to train a fully supervised instance segmentation method.

Figure 3: **Inference.** At test time, only the trained Mask-RCNN is required to output the prediction masks in the image. As an optional refinement step, the predicted masks can be replaced with the object proposals of highest Jaccard similarity.

3.2 Fully Supervised Segmentation Branch

We can construct the segmentation labels for all the training images by using the trained pseudo mask generation branch. These are used as supervision to train a Mask R-CNN (shown as green components in Figure 2). Depending on the application, other choices of fully supervised methods can be used instead of Mask R-CNN: if the goal is to perform instance segmentation at real-time, one can consider training a YOLACT [6], and for semantic segmentation, one can consider training a DeepLab [8] segmentation network.

Mask R-CNN [16] combines Faster R-CNN [37] and FCN-based [27] methods to first detect the objects and then segment them. For an image I, with target pseudo masks T, Mask R-CNN with parameters θ is trained by optimizing the following objective function,

$$\mathcal{L}(I,T,\theta) = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{box}} + \mathcal{L}_{\text{mask}}, \tag{2}$$

where $\mathcal{L}_{\text{cls}}$ is the classification loss for the detected objects, $\mathcal{L}_{\text{box}}$ is the localization loss for the detected objects, $\mathcal{L}_{\text{mask}}$ is their segmentation loss. These terms are explained in more detail in the original Mask R-CNN paper [16].

At test time, we can simply use the trained Mask R-CNN to predict the object masks for an unseen image. To refine these masks, we leverage the same object proposal method as that used in training. In turn, we replace each predicted object mask with the proposal of highest Jaccard similarity. Figure 3 illustrates how this refinement process can lead to a better object mask.

4 Experiments

In this section, we demonstrate the efficacy of our method by comparing it against previous methods and analyzing the pseudo masks.

4.1 Experimental Setup

We follow the setup by [9, 44] for a fair benchmark, where the model only has access to an off-the-shelf proposal method and image-level labels for the training set. Also from their setup, we adopt the evaluation metric, mean average precision for Intersection-over-Union (IoU) of 0.25, 0.5, and 0.75.

Like other works in the literature of weakly supervised methods, we perform all comparisons on the PASCAL VOC 2012 dataset [13]. The dataset represents a diverse set of everyday scenes. It is divided into 1442 images for training, and 1449 images for validation.

Method	Supervision	mAP25	mAP50	mAP75	ABO
Mask R-CNN [16]	pixel-level	58.9	51.4	32.4	-
DeepMask [28]	pixel-level	-	41.7	09.7	-
PRM [44]	image-level	44.3	26.8	09.0	37.6
PRM+Density [9]	image-level++	48.5	30.2	14.4	44.3
DeepMask [28]	bounding box	39.4	08.1	-	-
WISE (Ours)	image-level	48.5	40.4	22.2	51.3
WISE+Refine (Ours)	image-level	**49.2**	**41.7**	**23.7**	**55.2**

Table 1: **PASCAL VOC 2012.** Comparison of our framework (WISE) against other methods on various levels of supervision. WISE+Refine uses the refinement step shown in Figure 3. Mask R-CNN and DeepMask use full supervision, whereas PRM uses image-level labels. Same as WISE, PRM and PRM+Density leverage a pretrained proposal method. Requiring stronger supervision than WISE, DeepMask and PRM+Density have access to bounding box and image-level counts, respectively.

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	Avg.
Mask R-CNN	71.2	0.3	72.2	53.2	29.8	68.7	47.3	77.1	13.3	54.7	41.0	65.5	51.5	69.6	57.8	31.0	46.9	45.6	69.7	61.4	51.4
WISE	59.2	0.6	62.6	38.6	18.8	57.3	31.7	66.9	8.3	40.5	11.0	55.5	48.7	60.2	34.4	24.4	38.3	33.1	61.7	56.9	40.4
WISE+Refine	63.2	0.3	60.7	39.1	21.0	59.4	31.9	68.6	9.2	43.1	15.6	58.0	48.6	62.3	36.4	21.9	38.8	34.3	65.5	56.9	41.7

Table 2: **PASCAL VOC 2012.** Per-class comparison against the mAP_{50} metric on PASCAL VOC 2012 validation set. Mask R-CNN was trained with the ground-truth per-pixel labels.

Annotators for this dataset acquired per-pixel labels for 20 objects, ranging from inanimate objects such as airplanes and bikes, and living objects such as humans and horses. However, we only use the image-level labels to train our methods.

4.2 Implementation Details

We discuss our method's procedure and parameters below. We also plan to make the code publicly available.

Network architecture. As a common practice, we use the ResNet-50 [15] that is pretrained on ImageNet [12] as the backbone for PRM and Mask R-CNN. Unlike PRM, Mask R-CNN's backbone is equipped with a feature pyramid network [26] that extracts features at different resolutions. The pretrained weights, along with the rest of the parameters, are then finetuned on the PASCAL VOC 2012 training set. The remaining parameters of PRM and Mask R-CNN are in the implementation details discussed in Zhou et al. [44], and He et al. [16], respectively. We used a pretrained SharpMask [32] for our proposal method.

Optimization parameters. Following the official code of Mask R-CNN, we scale its input images so that the short axis has a minimum of 800px and the long axis a maximum of 1333px. Using a single GPU of TitanX, we set our batch size as 1 and train using the SGD optimizer with a learning rate of 0.00125 for 50K iterations. This learning rate was adjusted from He et al. [16], where they used a bigger batch size. We also augment the dataset with horizontal flips and color jittering as recommended by Deng et al. [12]. PRM was trained as described in Zhou et al. [44].

Metric	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	Avg.
Mask R-CNN + GT	92.4	15.1	97.4	87.9	91.4	94.4	93.8	100	68.2	93.4	88.8	97.4	96.4	95.3	92.8	89.3	92.3	97.7	100	100	89.2
Pseudo Masks	24.5	1.0	29.1	18.7	11.3	38.6	26.6	43.1	8.0	35.6	6.1	38.8	46.2	23.8	10.7	7.4	35.9	29.4	41.6	39.1	25.8
WISE	43.8	3.2	43.8	35.9	16.8	51.9	36.3	56.8	7.3	45.8	15.1	53.5	59.8	45.5	18.2	10.9	47.3	38.9	61.5	58.5	37.5

Table 3: **PASCAL VOC 2012 training set.** Comparison of the generated pseudo masks, and WISE's predicted masks with respect to mAP50. WISE was trained on a set of pseudo masks, and was able to output better masks for the same training images. Mask R-CNN + GT was trained on the ground-truth per-pixel labels.

4.3 Comparison to Previous Work

We first quantitatively compare our approach against previous methods that use the same supervision as ours; that is, image-level labels, an object proposal method, and a ResNet-50 backbone pretrained on ImageNet. Table 1 summarizes the results on the PASCAL VOC 2012 dataset. Our method significantly outperforms the current state-of-the-art by a large margin with respect to Average Best Overlap (ABO) [34], mAP25, mAP50, and mAP75. Further, WISE without refinement also beats current state-of-the-art. Even more so, our method outperforms Cholakkal et al. [9] which uses slightly stronger labels than image-level. Their labels distinguish between images with 0, 1, 2-4, and 4-or more objects. Figure 4 visualizes qualitative results of WISE for each category. We further report the per-class results in Table 2 and compare it against Mask R-CNN trained on the true masks with respect to mAP50. This illustrates that our results are competitive against fully-supervised methods.

Our method can also compete with those that use stronger supervision. Against DeepMask [18], our method outperforms two of their methods, one that uses bounding boxes as labels, and the other that uses full supervision as labels (see Table 1). Compared to Mask R-CNN trained on the pixel-level labels, our method still has a large room for improvement, which can be bridged by either improving the object localization component or the object proposal method.

While the overall results suggests that Mask R-CNN can effectively train from noisy, incomplete labels. The labels are noisy because the proposal masks are not perfect, and incomplete because PRM does not locate all the objects in the image. Indeed, we hypothesize that using a better object localizer such as that of Cholakkal et al. [9] would lead to better results. But we leave that for future work.

4.3.1 Analysis of Pseudo masks

We measure the generated pseudo masks performance by computing the mAP50 between the ground-truth and the generated masks. We also compute the mean absolute error to determine the number of identified objects in the images. These results are summarized in Table 3, which show that a large room for improvement is required for both metrics. Examples of the synthesized masks are shown in Figure 1, where one can see that the pseudo masks are not of high quality, yet the trained Mask R-CNN is able to output good masks in Figure 4.

4.3.2 Ablation Studies

The object sizes and the number of objects in an image can have severe impact on the performance of an instance segmentation model. Figure 5 shows that WISE struggles with segmenting small objects, and when the number of objects is larger than 4. In fact, there is a

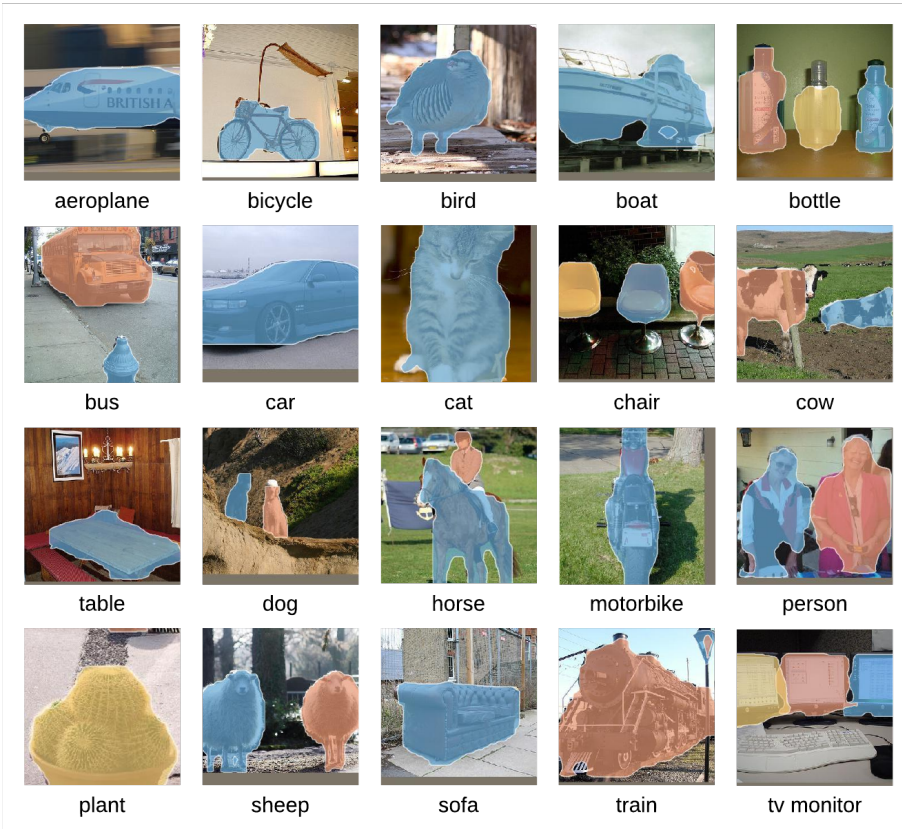

Figure 4: **Qualitative results.** Qualitative results of WISE on PASCAL VOC 2012 val. set. The images illustrate the predicted masks of the trained Mask R-CNN for different classes.

heavy decline in performance when the number of objects is more than 1. More robust than WISE, a Mask R-CNN trained on per-pixel labels is able to maintain higher performance with small objects and with images with larger number of objects. In addition, such Mask R-CNN performs significantly better than WISE for small objects. This suggests that the pseudo masks trained by WISE are likely far from accurate.

5 Conclusion

We proposed a weakly supervised instance segmentation method that follows a two-stage pipeline for training on image-level labels. In the first stage, it uses class activation maps with a peak stimulation layer to locate the objects in the training images, and then object proposals to generate pseudo masks for these objects. In the second stage, we use Mask R-CNN to train on the pseudo masks in a fully supervised manner. We evaluate on PASCAL VOC 2012, a standard benchmark for weakly supervised methods, where Mask R-CNN trained on pseudo masks outperformed not only methods with the same level of supervision, image-level labels, but also methods that use counts and bounding boxes in their supervision.

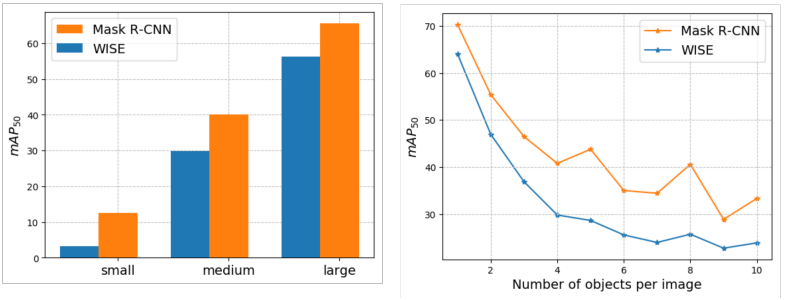

Figure 5: **Statistical Analysis.** The left figure illustrates the performance of WISE and a Mask R-CNN trained on per-pixel labels across various object sizes; and the right figure illustrates the same benchmark but across images with different number of objects.

References

[1] Jiwoon Ahn and Suha Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. *CVPR*, 2018.

[2] Jiwoon Ahn, Sunghyun Cho, and Suha Kwak. Weakly supervised learning of instance segmentation with inter-pixel relations. *ArXiv*, abs/1904.05044, 2019.

[3] Pablo Arbeláez, Jordi Pont-Tuset, Jonathan T Barron, Ferran Marques, and Jitendra Malik. Multiscale combinatorial grouping. In *CVPR*, 2014.

[4] Amy Bearman, Olga Russakovsky, Vittorio Ferrari, and Li Fei-Fei. WhatâĂŹs the point: Semantic segmentation with point supervision. In *ECCV*, 2016.

[5] Hakan Bilen and Andrea Vedaldi. Weakly supervised deep detection networks. In *CVPR*, 2016.

[6] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. Yolact: Real-time instance segmentation. *arXiv preprint arXiv:1904.02689*, 2019.

[7] Liang-Chieh Chen, Alexander Hermans, George Papandreou, Florian Schroff, Peng Wang, and Hartwig Adam. Masklab: Instance segmentation by refining object detection with semantic and direction features. In *CVPR*, 2018.

[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. *PAMI*, 40(4):834–848, 2018.

[9] Hisham Cholakkal, Guolei Sun, Fahad Shahbaz Khan, and Ling Shao. Object counting and instance segmentation with image-level supervision. In *CVPR*, 2019.

[10] 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*, 2016.

[11] Jifeng Dai, Kaiming He, and Jian Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. *CoRR*, 2015.

[12] 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.

[13] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 2010.

[14] Cheng-Yang Fu, Mykhailo Shvets, and Alexander C. Berg. RetinaMask: Learning to predict masks improves state-of-the-art single-shot detection for free. In *arXiv preprint arXiv:1901.03353*, 2019.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[16] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017.

[17] Jan Hosang, Rodrigo Benenson, Piotr Dollár, and Bernt Schiele. What makes for effective detection proposals? *T-PAMI*, 2016.

[18] Anna Khoreva, Rodrigo Benenson, Jan Hendrik Hosang, Matthias Hein, and Bernt Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In *CVPR*, 2017.

[19] Alexander Kolesnikov and Christoph H Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *ECCV*, 2016.

[20] Tomasz K. Konopczynski, Thorben Kröger, Lei Zheng, and Jürgen Hesser. Instance segmentation of fibers from low resolution ct scans via 3d deep embedding learning. In *BMVC*, 2018.

[21] Maksim Lapin, Matthias Hein, and Bernt Schiele. Analysis and optimization of loss functions for multiclass, top-k, and multilabel classification. *PAMI*, 40(7):1533–1554, 2018.

[22] Issam H Laradji, Negar Rostamzadeh, Pedro O Pinheiro, David Vazquez, and Mark Schmidt. Where are the blobs: Counting by localization with point supervision. In *ECCV*, 2018.

[23] Issam H. Laradji, Negar Rostamzadeh, Pedro O. Pinheiro, David Vázquez, and Mark W. Schmidt. Instance segmentation with point supervision. *ArXiv*, abs/1906.06392, 2019.

[24] Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *CVPR*, 2016.

[25] 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.

[26] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, 2016.

[27] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.

[28] Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Pablo Arbeláez, and Luc Van Gool. Convolutional oriented boundaries. In *ECCV*, 2016.

[29] 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. Accessed: March 12th 2019.

[30] Pedro O Pinheiro and Ronan Collobert. From image-level to pixel-level labeling with convolutional networks. In *CVPR*, 2015.

[31] Pedro O Pinheiro, Ronan Collobert, and Piotr Dollár. Learning to segment object candidates. In *NIPS*, 2015.

[32] Pedro O Pinheiro, Tsung-Yi Lin, Ronan Collobert, and Piotr Dollár. Learning to refine object segments. In *ECCV*, 2016.

[33] Regina Pohle and Klaus D Toennies. Segmentation of medical images using adaptive region growing. In *MIIP*, 2001.

[34] Jordi Pont-Tuset and Luc Van Gool. Boosting object proposals: From pascal to coco. In *CVPR*, 2015.

[35] Xiaojuan Qi, Zhengzhe Liu, Jianping Shi, Hengshuang Zhao, and Jiaya Jia. Augmented feedback in semantic segmentation under image level supervision. In *ECCV*, 2016.

[36] Mengye Ren and Richard S Zemel. End-to-end instance segmentation with recurrent attention. In *CVPR*, 2017.

[37] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015.

[38] Bernardino Romera-Paredes and Philip Hilaire Sean Torr. Recurrent instance segmentation. In *ECCV*, 2016.

[39] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 2017.

[40] Peng Tang, Xinggang Wang, Xiang Bai, and Wenyu Liu. Multiple instance detection network with online instance classifier refinement. In *CVPR*, 2017.

[41] Peng Tang, Xinggang Wang, Angtian Wang, Yongluan Yan, Wenyu Liu, Junzhou Huang, and Alan Yuille. Weakly supervised region proposal network and object detection. In *ECCV*, 2018.

[42] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. In *ICCV*, 2013.

[43] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016.

[44] Yanzhao Zhou, Yi Zhu, Qixiang Ye, Qiang Qiu, and Jianbin Jiao. Weakly supervised instance segmentation using class peak response. In *CVPR*, 2018.

[45] C Lawrence Zitnick and Piotr Dollár. Edge boxes: Locating object proposals from edges. In *ECCV*, 2014.