

Prior to Segment: Foreground Cues for Weakly Annotated Classes in Partially Supervised Instance Segmentation

David Biertimpel^{1,2*} Sindi Shkodrani² Anil S. Baslamisli¹ Nóra Baka²

¹University of Amsterdam ²TomTom

david.biertimpel@protonmail.com a.s.baslamisli@uva.nl
{sindi.shkodrani, nora.baka}@tomtom.com

Abstract

Instance segmentation methods require large datasets with expensive and thus limited instance-level mask labels. Partially supervised instance segmentation aims to improve mask prediction with limited mask labels by utilizing the more abundant weak box labels. In this work, we show that a class agnostic mask head, commonly used in partially supervised instance segmentation, has difficulties learning a general concept of foreground for the weakly annotated classes using box supervision only. To resolve this problem, we introduce an object mask prior (OMP) that provides the mask head with the general concept of foreground implicitly learned by the box classification head under the supervision of all classes. This helps the class agnostic mask head to focus on the primary object in a region of interest (RoI) and improves generalization to the weakly annotated classes. We test our approach on the COCO dataset using different splits of strongly and weakly supervised classes. Our approach significantly improves over the Mask R-CNN baseline and obtains competitive performance with the state-of-the-art, while offering a much simpler architecture.¹

1. Introduction

Instance segmentation is an essential task in computer vision with applications ranging from autonomous vehicles to robotics and medical imaging [5, 6, 13, 22, 26, 30]. A major contributor to the success of recent instance segmentation methods is the availability of large-scale datasets with numerous instance-level mask labels [7, 8, 12, 24, 36]. A major problem with mask labels is that their acquisition is rather time-consuming at ~ 67 seconds per instance [4]. Conversely, weak labels (w.r.t. instance masks) such as bounding boxes (~ 10.5 seconds) allow for much more

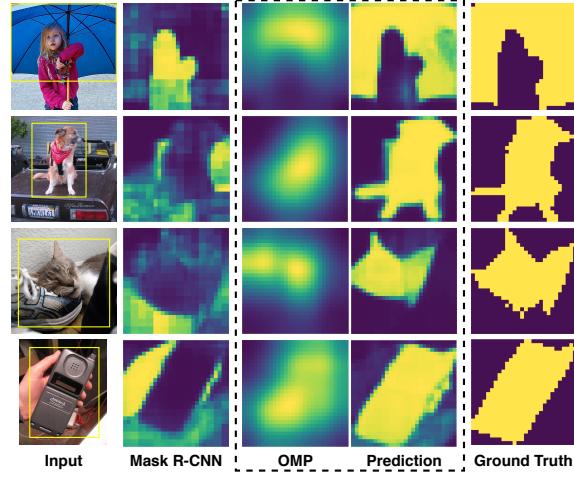


Figure 1. Our object mask prior (OMP) provides foreground cues to the mask head highlighting the primary instance in ambiguous RoIs (yellow pixels indicate foreground). OPMask is able to resolve ambiguous constellations and segment the correct instance, while our Mask R-CNN baseline fails to do so.

efficient annotation [4]. Nonetheless, for classes without sufficient mask labels, conventional instance segmentation methods perform poorly and tend to generate mask predictions that are perforated, do not cover the entire object or are completely missing it [16, 20]. To improve mask predictions for classes with no mask labels available, recent research has focused on addressing the problem in a partially supervised learning setting [9, 16, 20, 40], where all classes are annotated with box labels while only a subset of these classes also carry instance mask labels. The goal in partially supervised instance segmentation is to use the abundant but weak box labels in conjunction with the strong but limited mask labels to predict better instance masks for the weakly (box) annotated classes.

In current methods, the task of generalizing mask predictions for weak classes is either achieved with meta-learning

*This paper is the product of work during an internship at TomTom.

¹Code is available at: <https://github.com/dbtmp1/OPMask>

of class aware weights [16] or with a class agnostic mask head [9, 20, 40]. In the latter case, instead of predicting a mask per class, each pixel in the RoI is classified into either foreground or background. Thus, the class agnostic mask head faces the challenge of having to learn a general concept of foreground in order to generalize to unseen object classes. However, this often fails, even if abundant box labels are provided for the weak classes.

In this paper, we identify that the problem originates, on the one hand, from the ambiguous constellations between object instances, where pixels of one instance appear in the bounding box of the other. Thus, the actual foreground becomes ambiguous to the mask head when the RoI contains multiple and possibly overlapping instances. See Figure 1 for examples. On the other hand, instances of weak classes that appear in the background of a RoI during training are actively learned as background. This hurts generalization to weak classes that frequently interact with other supervised classes. To address these problems, we introduce an object mask prior (OMP) that highlights the correct foreground in each RoI. This helps the mask head to resolve ambiguous constellations, learn a more general concept of the foreground, and generalize it to weak classes.

Recent works have demonstrated that shape priors are beneficial inductive biases that steer models towards more stable mask predictions. For example, ShapeMask [20] creates a knowledge base of shape priors by applying k-means to the ground-truth masks, whereas ShapeProp [40] creates priors by using pixel-wise multiple instance learning on bounding boxes. Although these priors help to generalize to weak classes, they do not explicitly address the problems mentioned above.

Conversely, our prior is explicitly optimized to highlight the foreground in a RoI using the box supervision from all classes. This is achieved by exploiting the fact that the box classification head naturally learns to identify the primary class in a RoI. As the box head receives labels for all classes in the partially supervised setting, the box features capture a general concept of foreground. To reveal this foreground, we use class activation maps (CAMs) [39], which are coarse localization maps indicating the most discriminative image regions detected by the model. Therefore, given a correct classification, CAMs are expected to highlight foreground areas corresponding to the primary RoI class.

Unlike previous methods that introduce separate modules for prior creation, we only rely on the features in the box head and embed the OMP into our model in an end-to-end manner. Besides using box supervision from all classes, our prior is able to use mask gradients from the limited mask labels to increase its spatial extent. We embed our OMP in the Mask R-CNN meta architecture and name our overall model OPMask (Object Prior Mask R-CNN). Our main contributions are the following:

- We identify two fundamental problems in partially supervised instance segmentation: First, instances of weak classes appearing in the background of a mask supervised RoI during training are learned as background by the model. Second, in ambiguous RoIs containing multiple and possibly overlapping instances, the mask head has difficulties finding the foreground.
- We introduce an object mask prior (OMP) in the mask head to solve the above identified problems. The prior highlights the foreground across all classes by leveraging the information from the box head.
- On the COCO dataset [24], OPMask significantly improves over our Mask R-CNN baseline by 13.0 AP. Compared with the prior state-of-the-art, we improve over ShapeMask [20] and ShapeProp [40] and achieve competitive results against CPMask [9] while using a simpler architecture.

2. Related Work

Instance segmentation aims to segment every object instance in a scene. Detection based approaches [6, 13, 22, 26], which add a mask prediction network to existing detection models, represent the current state-of-the-art. Mask R-CNN [13] extends the two stage detection network Faster R-CNN [32] being the first to introduce a multi-task loss combining detection and mask gradients. Mask R-CNN is a strong baseline and often used as a meta-architecture due to its extensibility. Contour based approaches [27, 30, 38] segment objects by refining a sequence of vertices to match the object shape. Bottom-up approaches group pixels to generate instance masks [2, 25, 28]. As these approaches need large datasets with pixel-wise supervision, they are not suited for the partially supervised task.

Partially supervised instance segmentation. In partially supervised instance segmentation, a subset of classes is strongly annotated with box and mask supervision, while the remaining classes carry only weak box labels. The goal is to use the abundant box labels in conjunction with the limited masks to predict instance masks for all classes.

The pioneering approach by Hu *et al.* [16] augments a Mask R-CNN with a weight transfer function that learns a mapping from box to mask weights, introducing a class aware mask head capturing a representation for all classes.

Kuo *et al.* introduce ShapeMask [20] that creates a knowledge base of shape priors by applying k-means to the available ground-truth masks. A box embedding gives rise to a linear combination of the k-means centroids generating a shape prior that is further refined into an instance mask. ShapeMask bases its prior solely on the limited mask labels. In contrast, we use box labels of all available classes and use mask labels for refinement.

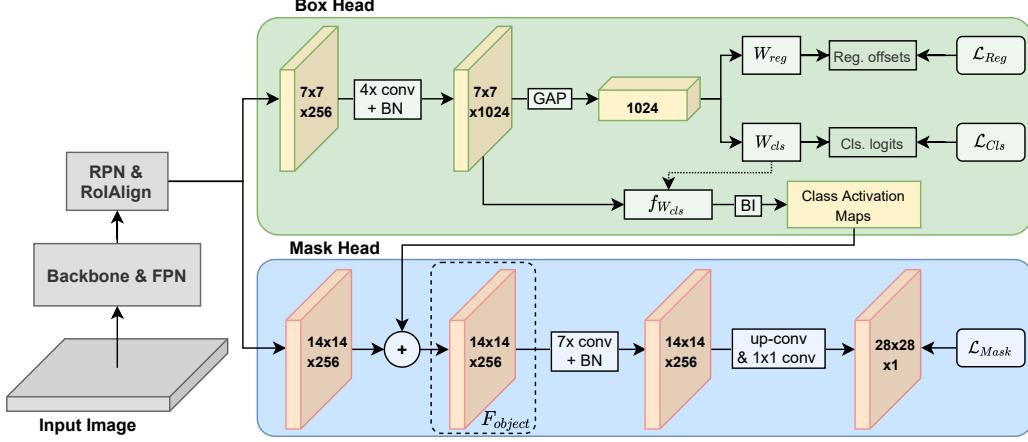


Figure 2. Overall architecture. The box head generates our OMP which is added to features entering the mask head to create object aware features F_{object} . The mask head then uses F_{object} to predict instance masks.

ShapeProp [40] uses pixel-wise multiple instance learning (MIL) on bounding boxes to create a saliency heatmap, which is further processed leading to a more expressive shape activation map. Both ShapeProp and OPMask utilize box labels to generate a prior for mask prediction. ShapeProp introduces two separate modules to generate and refine their prior. On the other hand, we take advantage of the fact that the box head *implicitly* learns a concept of foreground. Thus, we design our model to leverage the features that are already made available by the box head.

Finally, Fan *et al.* [9] learn the underlying shape and appearance commonalities between instance masks that should generalize to weak classes. The shape commonalities are learned by a boundary prediction head, while the appearance commonalities are enforced by an attention based affinity parsing module. Besides learning commonalities that aid generalization, we also identify that a major problem lies in ambiguous RoIs and the mask head having difficulties to learn a general concept of foreground. To address this, we utilize our OMP, which highlights the foreground of a ROI to resolve ambiguous constellations and help generalize to weak classes.

Weakly supervised instance segmentation approaches solely rely on weak labels such as bounding boxes or images level labels [1, 3, 10, 15, 19, 21, 41, 42]. Models using image-level labels [1, 10, 21, 41, 42] mostly use CAM based image-wide localization information to assist instance mask prediction. Zhou *et al.* [41] use the peaks of a class response map to detect and segment instances. Ge *et al.* [10] refine object attention maps using multi-task network heads sharing the same backbone. Both Laradji *et al.* [21] and Ahn *et al.* [1] create and refine pseudo masks which are later used to train a Mask R-CNN [13]. Setups where only image-level labels are available require the introduction of complex re-

finement modules. Conversely, in our setting, we rely on mask gradients that are already available in the model to improve our OMP.

Less work has been done using box supervision [15, 19]. Hsu *et al.* [15] employ a Mask R-CNN like architecture, where the mask head uses a MIL objective. Khoreva *et al.* [19] use GrabCut [33] to create pseudo ground truths to train a separate segmentation model. Instead of using box pixels to predict masks, we use CAMs to extract the foreground information in the box features to create our OMP.

3. Method

In partially supervised instance segmentation, a conventional Mask R-CNN with a class agnostic mask head fails to predict reliable instance masks for certain weak classes, as demonstrated in Figures 1 and 4. To address this, we propose OPMask which introduces an object mask prior (OMP) that captures foreground cues for all classes in the dataset (i.e. generalized foreground). OPMask follows the design of a Mask R-CNN [13] with a ResNet [14] backbone and FPN [23]. The model is illustrated in Figure 2.

3.1. Object Mask Prior (OMP)

The OMP functions as an inductive bias capturing a general concept of foreground to improve generalization to weak classes. In the partially supervised learning setup, predicting a general foreground is non-trivial for two main reasons: (1) pixel-wise mask labels are missing for a subset of classes, and (2) in many cases RoIs contain multiple and overlapping instances, making the foreground in a ROI ambiguous. The OMP tackles these issues by highlighting the correct foreground in each ROI, which helps the mask head to learn a more general concept of the foreground, resolve ambiguous ROIs, and generalize it to weak classes.

We create such a prior by extracting the foreground information captured by the box features in the box head. We use the fact that the box classification head learns a representation of the primary class (*i.e.* foreground) for all classes in the dataset. To reveal this foreground, we use class activation maps (CAMs) [39], which provide coarse localization maps emphasizing the most discriminative regions the model uses for classification. Consequently, given a correct classification, CAMs are expected to highlight foreground areas corresponding to the primary ROI class.

To realize CAM calculation, we use a box head with four convolution layers where Global Average Pooling (GAP) is applied on the last convolutional feature map. The resulting vector is processed by linear layers for box classification and regression (see Figure 2). We calculate CAMs with a function $f_{W_{cls}}$ which is a 1×1 convolution parameterized with the classification weights W_{cls} as follows:

$$M_{cam} = f_{W_{cls}}(F_{box}), \quad (1)$$

where F_{box} is the last feature map of the box head before GAP. This allows calculating all CAMs efficiently with a single operation while keeping them differentiable. Depending on whether it is training or inference time, we use the ground truth labels or the classes predicted by the box head to select the correct CAM slice from M_{cam} . We can expect CAMs to be stable across categories, since the partially supervised instance segmentation task assumes abundant box labels for all classes.

The CAMs of the correct class are added to the corresponding mask features as will be described in the next section. Apart from providing the mask head favorable foreground cues, this also allows the mask gradients to back-propagate through the box head. A well known shortcoming of CAMs is that they do not cover the full extent of the objects, but only the minimal area of the most distinctive features. Other works use expensive sampling or refinement steps to increase the spatial extent of CAMs [34, 41, 42]. To ensure a simple and fast architecture without separate CAM refinement steps, we use the efficient original CAM definition [39] and leverage the mask gradients to provide mask information to the features in the box head. This leads to an increase in the spatial extent of the CAMs allowing them to capture finer details.

As a result, CAMs that receive mask gradients give rise to our OMP. The fact that the OMP originates from the box classification task, which is directly optimized to classify the primary instance in a ROI, provides it with strong foreground cues. This is in direct contrast to previous work [20, 40] that introduce separate modules for prior creation and refinement. Conversely, our OMP is predestined to provide the mask head with a general concept of foreground allowing it to resolve ambiguous ROIs and also better generalize to weak classes.

3.2. Inducing the Prior

After generating the OMP, we aggregate it with the FPN features after ROIAlign F_{fpn} to create object-aware features F_{object} as follows:

$$F_{object} = F_{fpn} + M_{cam}, \quad (2)$$

where $M_{cam,k} \in \mathbb{R}^{H,W}$ is added to each channel of its matching ROI $F_{fpn,k} \in \mathbb{R}^{D,H,W}$. Before addition, we use bilinear interpolation to adjust M_{cam} to the spatial dimensions of F_{fpn} .

The addition highlights the features in F_{fpn} at the spatial locations where the response of the OMP is high. This provides the mask head with explicit foreground information that are embedded by the subsequent convolutional layers in the mask head. This incentivizes the mask head to learn a general concept of foreground for all classes in the dataset. Note that this approach is different from feature fusion between box and mask features as proposed in [17] and [31]. While features are a rich description of the image and its content, our OMP is a simple one layer object/background encoding helping the mask head focus on the correct object.

After the addition, F_{object} is processed by a function f_{mask} consisting of seven 3×3 convolution layers followed by one transposed convolution layer doubling the features spatial resolution and one 1×1 convolution performing mask prediction as follows:

$$M_{mask} = f_{mask}(F_{object}), \quad (3)$$

where M_{mask} is the mask prediction after applying a pixel-wise sigmoid. We use seven convolution layers to achieve a receptive field large enough such that f_{mask} operates on the entire input feature map. Batch normalization [18] is applied after each 3×3 convolution to utilize its stochastic properties to improve generalization. Finally, a pixel-wise binary cross-entropy loss is applied to M_{mask} using the available mask labels M_{gt} as follows:

$$\mathcal{L}_{Mask} = BCE(M_{mask}, M_{gt}). \quad (4)$$

4. Experiments

In Section 4.1, first, the dataset and experimental setup are introduced. Then, in Section 4.2, we provide evidence that instances of weak classes appearing in the background of a ROI during training are learned as background, and a conventional class agnostic mask head has difficulties considering the correct foreground in ambiguous ROIs. Afterwards, Section 4.3 shows the capabilities of OPMask to generalize to weak classes. Finally, in Section 4.4, we compare our OMP against regular CAMs showing the positive impact of mask gradients updating box features.

4.1. Experimental Setup

We conduct our experiments on the COCO dataset [24]. To realize the partially supervised learning setup, we split the *mask labels* of the 80 COCO *thing* classes into two subsets. One subset consists of the strongly annotated classes used for training, the other subset consists of the weakly annotated classes used for evaluation and vice versa. Box labels are available for all classes during training. To compare against related work, we mainly focus on dividing the COCO dataset between the 20 classes of the Pascal voc dataset [8] (inside COCO), and the remaining 60 unique COCO classes. In the following, non-voc \rightarrow voc denotes that the mask head is trained on the non-voc classes and evaluated on voc classes, vice versa the same applies for voc \rightarrow non-voc. For training, we use SGD with Momentum with an initial learning rate of 0.02 which is linearly warmed up for the first 1000 iterations [11]. The batch size is set to 16 and the gradients are stabilized by clipping them at a value of 1.0. ResNet-50 and ResNet-101 [14] with a FPN [23] are used as backbones. The implementation is based on PyTorch [29] and Detectron2 [37].

Baseline. We use a Mask R-CNN with a class agnostic mask head. For a fair comparison, we use the same box head as OPMask and also add batch norm to its mask head. In the following, we call this baseline ‘Our Mask R-CNN’.

4.2. Insights on Identifying Foreground in RoIs

Learning classes as background. A class agnostic mask head faces the task of classifying ROI pixels between foreground or background, where pixels that correspond to supervised classes are considered foreground, while all other pixels are regarded as background. The COCO dataset contains complex scenes with cluttered objects, which causes ROIs to often contain more than one instance. Background pixels can either be part of the available supervised classes, belong to weak classes to which we want to generalize, or not be part of any class in the dataset. In the second case, we face the dilemma that the model actively learns to classify features that correspond to weak classes as background. This clearly conflicts with the generalization goal of the partially supervised learning task.

This phenomenon particularly affects classes that frequently interact with other classes and thus appear more often in the background of a mask supervised ROI. To investigate this, we compute the correlation between class overlap and mask AP for weak classes (in voc \rightarrow non-voc and non-voc \rightarrow voc). To approximate the overlap between classes, we compute the IoU of all ground-truth bounding boxes in the COCO dataset. Afterwards, we compute a regression between the mean IoU of each class and its mask AP.

Two regression models are presented in Figure 3. The first (*left*) is computed with our Mask R-CNN baseline showing a significant negative correlation between mean

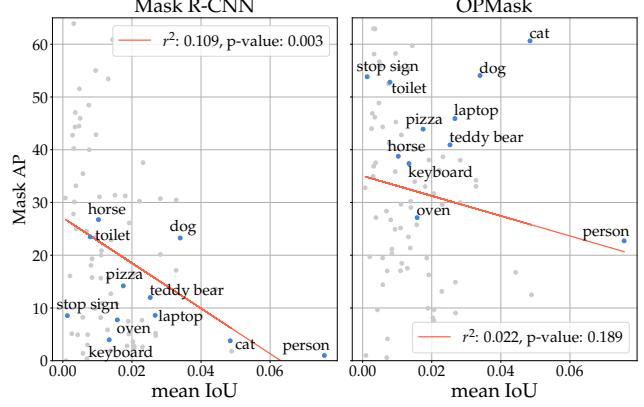


Figure 3. Regressions showing the correlation between box IoU and mask AP of all COCO classes. We compare mask AP scores of our Mask R-CNN baseline (*left*) and OPMask (*right*). The classes with the largest relative improvement are highlighted.

IoU and mask AP across all classes ($p = .003 < .01$). This provides evidence for our hypothesis that weak classes appearing in the background of ROIs are actively learned as background during training. The second regression (*right*), computed with OPMask, shows only a weak negative correlation that is not strong enough to reach significance ($p = .189 \not< .01$). At the same time, we see notable improvements for classes with high mean IoU values, which are more likely to appear in the background of other classes ROIs (*e.g.* *person*: 0.99 to 22.72 AP, *cat*: 3.77 to 60.63 AP). This suggests that the OMP is able to provide the mask head with a general concept of foreground, which counteracts learning these weak classes as background.

Resolving ambiguous ROIs. Another issue with multiple and possibly overlapping instances is that the primary instance (*i.e.* foreground) of the ROI may be ambiguous. We identify that a conventional mask head has difficulties to locate the foreground in these ambiguous constellations.

Model	Full COCO		non-voc \rightarrow voc		voc \rightarrow non-voc	
	Amb.	\neg Amb.	Amb.	\neg Amb.	Amb.	\neg Amb.
Our Mask R-CNN	15.9	36.2	10.8	27.4	6.5	19.7
OPMask	20.5	38.0	19.5	37.2	17.1	32.2

Table 1. Mask AP of OPMask vs. our Mask R-CNN baseline in ambiguous and non-ambiguous instances. OPMask shows better performance in all comparisons, with the largest improvements in ambiguous constellations in the partially supervised setup.

To quantify the effect of ambiguous instances on model performance, we split the COCO validation set into ambiguous and non-ambiguous instances. Since properly quantifying ambiguity is non-trivial, we avoid complicated heuristics and use the IoU of the box labels as a simple proxy. This allows to capture general trends while putting emphasis on simplicity and reproducibility. We consider instances with a box label IoU ≥ 0.5 to at least one other instance in an

image as ambiguous. Table 1 compares OPMask with our Mask R-CNN baseline (both ResNet-50 backbones) in ambiguous and non-ambiguous instances, trained in the fully- and partially supervised COCO splits. The results show that OPMask performs better than our Mask R-CNN baseline in all comparisons, with the largest improvements in ambiguous constellations. Especially in the partially supervised setting, our Mask R-CNN baseline falls significantly behind OPMask. This suggests that a conventional class agnostic mask head has considerable problems with ambiguous instances, especially when it has to generalize to weak classes. At the same time, OPMask manages to significantly mitigate this drop in mask AP thanks to the OMP highlighting the correct foreground in the RoIs. These results particularly reflect the precision of our OMP, as it is able to focus the correct object even in such ambiguous constellations.

In addition, Figure 1 provides qualitative comparisons between OPMask and our Mask R-CNN baseline handling ambiguous instances. The results present ambiguous RoIs where our Mask R-CNN baseline falsely predicts background instances as foregrounds. On the other hand, the OMP is able to highlight the foreground instance in the ROI allowing OPMask to correctly segment the instance. All examples are from models trained either in the voc → non-voc or non-voc → voc setting. Interestingly, the results in the first row are achieved with models trained in the voc → non-voc setting, where person is a supervised class and umbrella is a weak class. While the Mask R-CNN incorrectly segments the person, OPMask identifies the umbrella as the primary class and is able to predict an accurate instance mask while generalizing to a weak class. This further emphasizes the quantitative results where our Mask R-CNN baseline performs particularly poorly for ambiguous instances in the partially supervised setting.

4.3. Generalization to Weakly Annotated Classes

Pascal voc vs. non-voc classes. We present the quantitative results for the voc vs. non-voc splits in Table 2. The results show that OPMask considerably improves over our Mask R-CNN baseline in all cases. For example, with a ResNet-50 backbone, a significant increase of 10.1 AP in non-voc → voc and 13.0 AP in voc → non-voc is achieved. OPMask also performs better than previous approaches ShapeProp [40] and ShapeMask [20] in all cases. It is notable that even with a ResNet-50, we achieve better or competitive performance than ShapeMask and ShapeProp that are equipped with the stronger ResNet-101. When comparing to the current state-of-the-art CPMask [9], we achieve substantial improvements with a ResNet-50 backbone (*e.g.* 3.1 AP). For the ResNet-50 backbone CPMask only provides results in the voc → non-voc setting. Our model with a ResNet-101 backbone achieves competitive performance in non-voc → voc (*e.g.* increase of 0.3 AP),

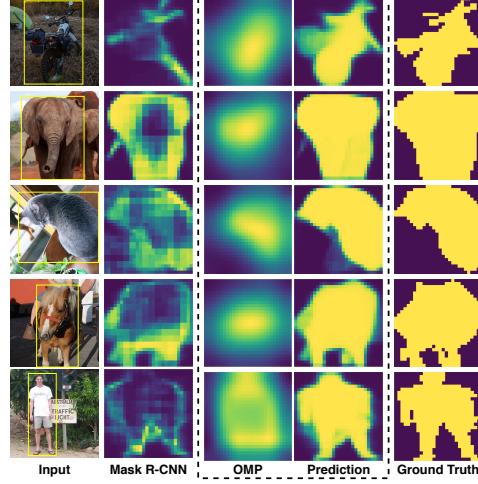


Figure 4. The Mask R-CNN baseline produces perforated, incomplete or missing masks. OPMask driven by the OMP is able to accurately segment each instance of a weak class.

but also slightly worse performance in voc → non-voc (*e.g.* 0.8 AP decrease). It should be noted, however, that unlike us, CPMask uses multi-scale training when equipped with a ResNet-101 backbone, which is known to substantially increase the overall performance of the model. The fact that we significantly outperform CPMask without multi-scale training on a ResNet-50, but only achieve competitive performance on a ResNet-101 demonstrates the magnitude of improvements possible with multi-scale training. In addition, CPMask adopts the more capable FCOS [35] for box detection, which yields much better detection performance than a Faster R-CNN (+5.3 AP), positively affecting instance segmentation performance. We also emphasize the computational efficiency of OPMask. At inference time, our method (.12 s/img) is much more efficient than ShapeMask (29.1 s/img) and ShapeProp (.22 s/img), and only slightly heavier than our Mask R-CNN baseline (.08 s/img)². The code of CPMask is not available yet. Nevertheless, CPMask uses an extra boundary-parsing head requiring additional boundary labels and a self-attention based affinity-parsing module that produces expensive $14^2 \times 14^2 \times 256$ feature maps. Further, both modules introduce an additional loss for which gradients must be computed. Thus, in terms of computing overhead, OPMask emerges as a simpler approach than our related work.

Qualitative Results.

In Figures 1 and 4, we provide qualitative insights into how the OMP steers mask predictions and improves generalization to weak classes. Each example shows a weak class in either the voc → non-voc or non-voc → voc setup. Next to the OMP and mask prediction of OPMask, our Mask R-

²All inference times were obtained with a Nvidia Tesla V100.

Backbone	Method	non-voc → voc: test on voc						voc → non-voc: test on non-voc					
		AP	AP ₅₀	AP ₇₅	AP _S	AP _M	AP _L	AP	AP ₅₀	AP ₇₅	AP _S	AP _M	AP _L
R-50-FPN	Mask R-CNN [13]	23.9	42.9	23.5	11.6	24.3	33.7	19.2	36.4	18.4	11.5	23.3	24.4
	Our Mask R-CNN	26.4	46.4	26.7	14.2	26.4	36.5	18.9	35.5	18.4	12.4	22.8	22.9
	Mask ^X R-CNN [16]	28.9	52.2	28.6	12.1	29.0	40.6	23.7	43.1	23.5	12.4	27.6	32.9
	Mask R-CNN w/ ShapeProp [40]	34.4	59.6	35.2	13.5	32.9	48.6	30.4	51.2	31.8	14.3	34.2	44.7
	CPMask [9]	-	-	-	-	-	-	28.8	46.1	30.6	12.4	33.1	43.4
R-101-FPN	OPMask	36.5	62.5	37.4	17.3	34.8	49.8	31.9	52.2	33.7	16.3	35.2	46.5
	Mask R-CNN [13]	24.7	43.5	24.9	11.4	25.7	35.1	18.5	34.8	18.1	11.3	23.4	21.7
	Our Mask R-CNN	27.7	48.0	28.2	13.6	28.6	38.0	21.0	39.2	20.5	13.5	26.4	23.9
	Mask ^X R-CNN [16]	29.5	52.4	29.7	13.4	30.2	41.0	23.8	42.9	23.5	12.7	28.1	33.5
	ShapeMask [20]	33.3	56.9	34.3	17.1	38.1	45.4	30.2	49.3	31.5	16.1	38.2	28.4
	Mask R-CNN w/ ShapeProp [40]	35.5	60.5	36.7	15.6	33.8	50.3	31.9	52.1	33.7	14.2	35.9	46.5
	CPMask [9]	36.8	60.5	38.6	17.6	37.1	51.5	34.0	53.7	36.5	18.5	38.9	47.4
	OPMask	37.1	62.5	38.4	16.9	36.0	50.5	33.2	53.5	35.2	17.2	37.1	46.9

Table 2. Comparing OPMask with the state-of-the-art in the partially supervised instance segmentation setup on COCO. OPMask outperforms our Mask R-CNN baseline as well as our related work Mask^X R-CNN, ShapeMask and ShapeProp with both backbones. Further, OPMask performs better than CPMask on a ResNet-50 and remains competitive on a ResNet-101. non-voc → voc denotes that the mask head is trained on the non-voc classes and evaluated on voc classes, vice versa the same applies for voc → non-voc.

CNN baseline predictions are presented. The results show that the OMPs properly identify and highlight the primary instances in the RoIs while covering most of the objects’ spatial extend. Furthermore, we realize that our coarse prior is sufficient to enable the mask head to generalize to a refined mask. This underlines our hypothesis that it is of particular importance to provide the class agnostic mask head with a general concept of foreground across all classes. Finally, Figure 5 presents a number of COCO images with overplayed mask predictions produced in the voc → non-voc setup. The results show OPMask’s ability to generate precise predictions for instances of weak classes across different scenarios and object sizes. All examples in this section are achieved with models equipped with a ResNet-101.

Strongly vs weakly supervised class ratios. To provide a better overview of OPMask’s generalization abilities, we evaluate its performance on different splits of strongly and weakly annotated classes. To create the 40 class split, we start with the 20 Pascal voc [8] classes and randomly add another 20 classes from the non-voc split. Figure 6 shows that OPMask consistently improves over our Mask R-CNN baseline while demonstrating stable performance across all class splits. Even in the fully supervised setup (Full coco), where a Mask R-CNN remains competitive with other state-of-the-art instance segmentation methods, OPMask achieves better performance (+1.6 AP). We attribute these improvements to the fact that even with full supervision, the OMP helps the class agnostic mask head resolve ambiguous RoIs, which in return improves OPMask predictions. This shows that our OMP remains beneficial when strong annotations are available for all classes.

4.4. Refining the Object Mask Prior

A simple CAM as the OMP might do a reasonable job, though a better prior is expected to lead to a better segmen-

tation result. To improve our OMP, we let mask gradients backpropagate through the box head, which augments the box features with mask information. This enhances the CAMs by increasing their spatial extent and allowing the OMP to cover larger parts of the objects. The resulting refinement further improves the final mask AP by 1.1 points in non-voc → voc with ResNet-50 backbone. To investigate the improvement of the prior, we compare the mask AP of the OMP with vanilla CAMs on the COCO validation set. We compare against a Faster R-CNN and a Mask R-CNN with the same box head as OPMask. In Table 3, AP and AP₅₀ results of voc vs. non-voc class splits are provided. Since the Faster R-CNN does not receive any mask gradients, it is only trained and evaluated on all classes.

Model	train set	test on all		test on voc		test on non-voc	
		AP	AP ₅₀	AP	AP ₅₀	AP	AP ₅₀
Faster R-CNN	all	0.2	1.0	0.3	2.5	0.1	0.6
Mask R-CNN	all	0.2	1.3	0.3	2.3	0.1	1.0
OPMask	all	8.8	34.1	9.9	40.4	8.4	32.0
Mask R-CNN	voc	0.3	1.9	0.6	4.0	0.2	1.0
OPMask	voc	5.0	21.8	9.9	38.5	3.3	16.3
Mask R-CNN	non-voc	0.2	1.5	0.4	2.8	0.2	1.0
OPMask	non-voc	8.0	31.5	7.0	32.2	8.3	31.2

Table 3. Quantitatively comparing our OMP with CAMs of a Faster R-CNN and Mask R-CNN. The results show that our OMP is able to cover larger parts of the objects than conventional CAMs.

The results show that the OMP is significantly better than the CAMs of Faster R-CNN and Mask R-CNN. This underlines the positive influence of mask gradients on box head features and consequently on the OMP. The low AP values of the CAMs generated by Faster R-CNN and Mask R-CNN are caused by the fact that they often do not surpass the pixel-wise IoU threshold (i.e. ≥ 0.5), and are mostly considered negatives. The Mask R-CNN, where the backbone features are augmented with mask gradients, does not show significant improvements over the Faster R-CNN. This sug-

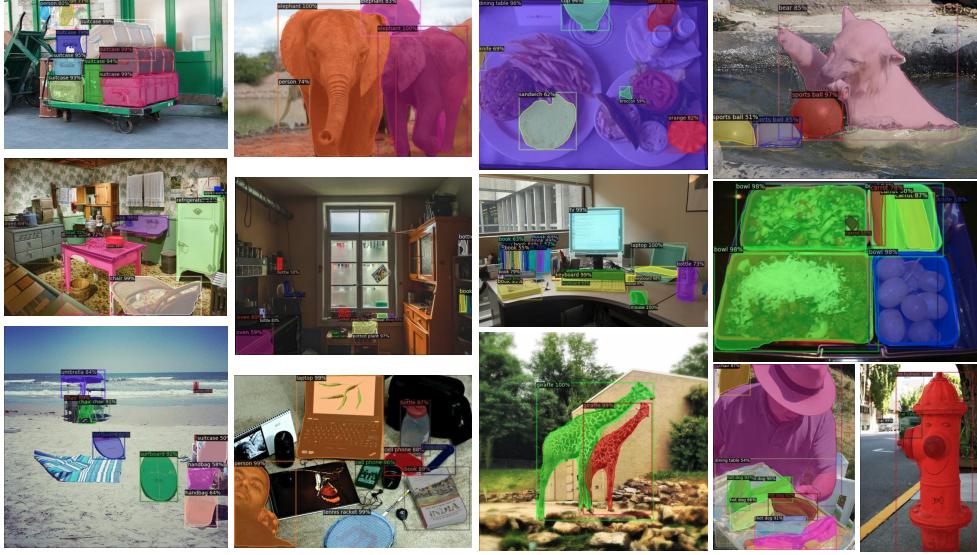


Figure 5. Qualitative results on COCO using the $\text{voc} \rightarrow \text{non-voc}$ split for training. This shows the ability of OPMask to predict precise instance masks for objects of weak classes across different scenes, and various object sizes and appearances.

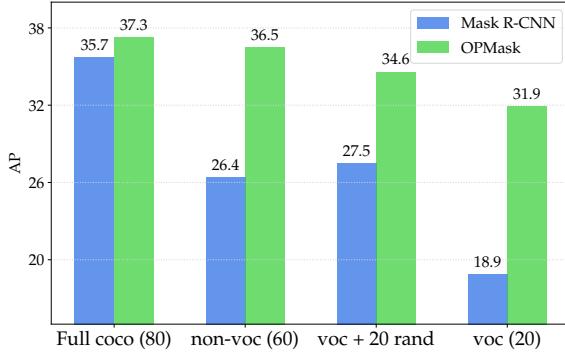


Figure 6. Performance of OPMask on varying numbers of supervised classes. It significantly improves over our Mask R-CNN baseline across all class splits, including the fully supervised setup.

gests that for CAM refinement, mask gradients should impact the features that are directly used to calculate the CAM activations. Finally, Figure 7 demonstrates qualitative improvements of CAMs on a number of COCO images.

5. Conclusion

We proposed OPMask, a novel approach to partially supervised instance segmentation. OPMask introduces an object mask prior (OMP) that helps its class agnostic mask head to learn a general concept of foreground, resolve ambiguous RoIs and generalize to weak classes. Our research pointed out two major problems hindering a class agnostic mask head to generalize to weak classes. First, instances of weak classes appearing in the background of a mask su-

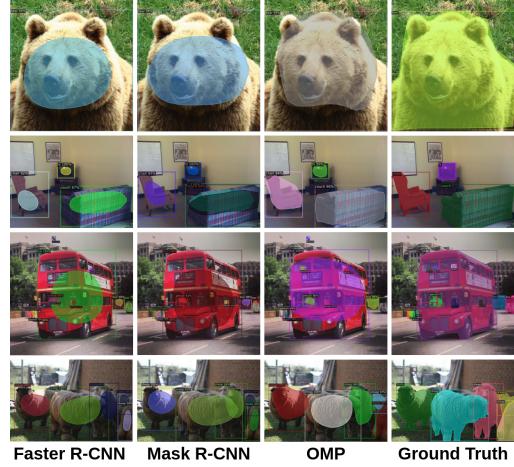


Figure 7. Comparing our OMP with CAMs from a Faster R-CNN and Mask R-CNN on COCO images. We see that our OMP is able to cover larger parts of the objects than regular CAMs.

pervised ROI during training are learned as background by the model. Second, in ambiguous RoIs that contain multiple and possibly overlapping instances, the mask head has difficulties to consider the correct foreground. We demonstrated that both problems can be vastly alleviated with our OMP that highlights foreground across all classes by leveraging the information from the box head. Finally, we showed that OPMask significantly improves over our Mask R-CNN baseline and achieves competitive performance with the state-of-the-art, while offering a much simpler architecture with less computational overhead.

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 (CVPR)*, June 2019.
- [2] Anurag Arnab and Philip HS Torr. Pixelwise instance segmentation with a dynamically instantiated network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 441–450, 2017.
- [3] Aditya Arun, C. V. 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 Max Welling, Nicu Sebe, Jiri Matas, and Bastian Leibe, editors, *Computer Vision - 14th European Conference, ECCV 2016, Proceedings*, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pages 549–565, Germany, 2016. Springer Verlag.
- [5] Hao Chen, Kunyang Sun, Zhi Tian, Chunhua Shen, Yongming Huang, and Youliang Yan. Blendmask: Top-down meets bottom-up for instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8573–8581, 2020.
- [6] Tianheng Cheng, Xinggang Wang, Lichao Huang, and Wenyu Liu. Boundary-preserving mask r-cnn. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 660–676, Cham, 2020. Springer International Publishing.
- [7] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3213–3223, 2016.
- [8] Mark Everingham, Luc Gool, Christopher K. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vision*, 88(2):303–338, June 2010.
- [9] Qi Fan, Lei Ke, Wenjie Pei, Chi-Keung Tang, and Yu-Wing Tai. Commonality-parsing network across shape and appearance for partially supervised instance segmentation. In *ECCV*, 2020.
- [10] Weifeng Ge, Sheng Guo, Weilin Huang, and Matthew R Scott. Label-penet: Sequential label propagation and enhancement networks for weakly supervised instance segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3345–3354, 2019.
- [11] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large mini-batch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.
- [12] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5356–5364, 2019.
- [13] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017.
- [14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [15] 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 H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32, pages 6586–6597. Curran Associates, Inc., 2019.
- [16] Ronghang Hu, Piotr Dollár, Kaiming He, Trevor Darrell, and Ross Girshick. Learning to segment every thing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4233–4241, 2018.
- [17] Zhaojin Huang, Lichao Huang, Yongchao Gong, Chang Huang, and Xinggang Wang. Mask scoring r-cnn. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [18] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. volume 37 of *Proceedings of Machine Learning Research*, pages 448–456, Lille, France, 07–09 Jul 2015. PMLR.
- [19] Anna Khoreva, Rodrigo Benenson, Jan Hosang, Matthias Hein, and Bernt Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.
- [20] Weicheng Kuo, Anelia Angelova, Jitendra Malik, and Tsung-Yi Lin. Shapemask: Learning to segment novel objects by refining shape priors. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.
- [21] Issam H Laradji, David Vazquez, and Mark Schmidt. Where are the masks: Instance segmentation with image-level supervision. *BMVC*, 2019.
- [22] Youngwan Lee and Jongyoul Park. Centermask: Real-time anchor-free instance segmentation. In *CVPR*, 2020.
- [23] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.
- [24] 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 David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing.
- [25] Shu Liu, Jiaya Jia, Sanja Fidler, and Raquel Urtasun. Sgn: Sequential grouping networks for instance segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3496–3504, 2017.

- [26] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [27] Diego Marcos, Devis Tuia, Benjamin Kellenberger, Lisa Zhang, Min Bai, Renjie Liao, and Raquel Urtasun. Learning deep structured active contours end-to-end. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8877–8885, 2018.
- [28] Davy Neven, Bert De Brabandere, Marc Proesmans, and Luc Van Gool. Instance segmentation by jointly optimizing spatial embeddings and clustering bandwidth. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8837–8845, 2019.
- [29] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.
- [30] Sida Peng, Wen Jiang, Huaijin Pi, Xiuli Li, Hujun Bao, and Xiaowei Zhou. Deep snake for real-time instance segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
- [31] Lu Qi, Li Jiang, Shu Liu, Xiaoyong Shen, and Jiaya Jia. Amodal instance segmentation with kins dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [32] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 91–99. Curran Associates, Inc., 2015.
- [33] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. Grabcut -interactive foreground extraction using iterated graph cuts. *ACM Transactions on Graphics (SIGGRAPH)*, August 2004.
- [34] Krishna Kumar Singh, Dhruv Mahajan, Kristen Grauman, Yong Jae Lee, Matt Feiszli, and Deepti Ghadiyaram. Don’t judge an object by its context: Learning to overcome contextual bias. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
- [35] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. FCOS: Fully convolutional one-stage object detection. In *Proc. Int. Conf. Computer Vision (ICCV)*, 2019.
- [36] Syed Waqas Zamir, Aditya Arora, Akshita Gupta, Salman Khan, Guolei Sun, Fahad Shahbaz Khan, Fan Zhu, Ling Shao, Gui-Song Xia, and Xiang Bai. isaid: A large-scale dataset for instance segmentation in aerial images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 28–37, 2019.
- [37] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. <https://github.com/facebookresearch/detectron2>, 2019.
- [38] Wenqiang Xu, Haiyang Wang, Fubo Qi, and Cewu Lu. Explicit shape encoding for real-time instance segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5168–5177, 2019.
- [39] B. Zhou, A. Khosla, Lapedriza. A., A. Oliva, and A. Torralba. Learning Deep Features for Discriminative Localization. *CVPR*, 2016.
- [40] Yanzhao Zhou, Xin Wang, Jianbin Jiao, Trevor Darrell, and Fisher Yu. Learning saliency propagation for semi-supervised instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10307–10316, 2020.
- [41] Yanzhao Zhou, Yi Zhu, Qixiang Ye, Qiang Qiu, and Jianbin Jiao. Weakly supervised instance segmentation using class peak response. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3791–3800, 2018.
- [42] Yi Zhu, Yanzhao Zhou, Huijuan Xu, Qixiang Ye, David Doermann, and Jianbin Jiao. Learning instance activation maps for weakly supervised instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3116–3125, 2019.