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

Issam H. Laradji - David Vazquez - Mark Schmidt

ELEMENT AI

THE UNIVERSITY OF BRITISH COLUMBIA

## Motivation

**Most Instance segmentation methods** require **very costly** human effort
- They need per-pixel labels

**Our method** requires **significantly less** human effort
- It only needs image-level labels

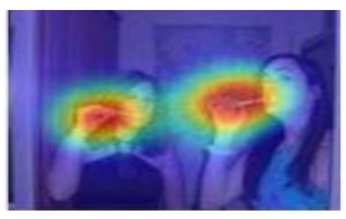
**Per-pixel labels**
(Mask R-CNN)

**Point-level annotations**

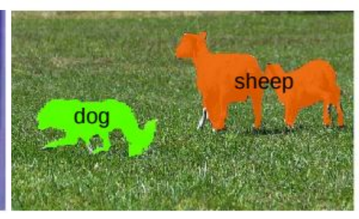

**Image-level annotation**
(Ours)

## Related work

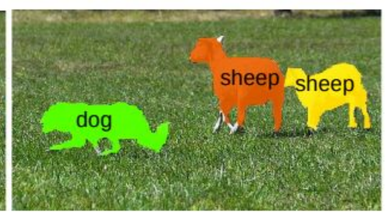
### Image-level labels as weak supervision



CAM: Class Activation Map
[Zhou et al. 2016]




PRM: Peak Response Map
[Zhou et al. 2018]




PRM + counting
[Cholakkal et al. 2019]

### Learning with pseudo labels

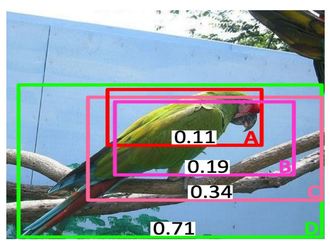
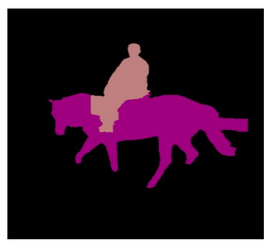



Image-level supervision
object detection
[Dai et al. 2015]

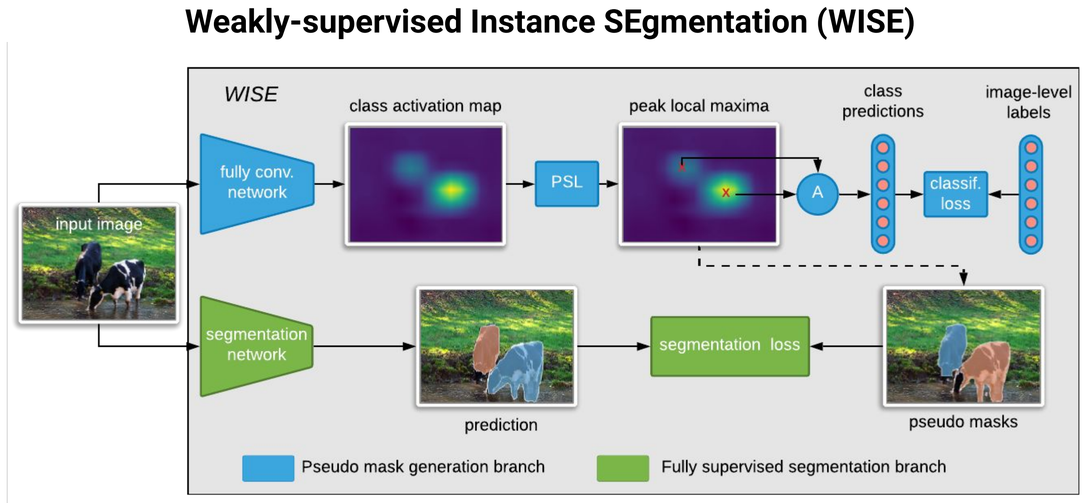


Bbox supervision
semantic segmentation
[Tang et al. 2017]

Image-level supervision
instance segmentation
(Ours)

## Proposed Method: WISE

### Weakly-supervised Instance SEgmentation (WISE)




### Step 1 - Pseudo Mask Generation Branch

**Localizing the objects (with image-level labels)**
- First, generate the class activation map to find the object regions
- Then, obtain the peak local maximas as object centroids
- Finally, take the average of the local maximas for multi-label classification

**Generating the training pseudo-labels**
- Generate 1000 proposal masks from a pretrained SharpMask
- Replace each predicted object location obtained with an proposal mask
  - De-noising strategy: select a proposal randomly based on its objectness

### Step 2 - Fully Supervised Segmentation Branch
- Using Step (1), construct the per-pixel labels for all training images
- Train a Mask R-CNN on these labels

### Prediction
- At test time only the trained Mask R-CNN is used
- Refinement: Replace each predicted object mask with the proposal of highest IoU


## Experimental setup

### Evaluation metric
- Mean average precision for Intersection-over-Union (IoU) of 0.25, 0.5, and 0.75.

### Dataset
- PASCAL VOC 2012 using only image-level annotations

### Implementation details

**Network architecture**
- Backbone: ResNet-50 ImageNet pretrained

**Optimization**
- Scale images to so that the short axis has a minimum of 800px and the long axis a maximum of 1333px
- Batch size 1
- SGD with learning rate of 0.00125 for 50K iterations
- Data augmentation: horizontal flips and color jittering

## Quantitative results

### Results on PASCAL VOC 2012 dataset

| Method | Supervision | mAP25 | mAP50 | mAP75 | ABO |
|---|---|---|---|---|---|
| Mask R-CNN [16] | pixel-level | 58.9 | 51.4 | 32.4 | - |
| DeepMask [18] | 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 [18] | 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** |

| 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 |

## Qualitative results


## Summary

### WISE
- Generate pseudo-masks using a PRM procedure and object proposals
- Train Mask-RCNN using these proposals

### Future Work
- Train using count-level supervision in order to extend it to crowded datasets

Scan me

## References

[He et al. 2017] - Mask R-CNN.
[Zhou et al. 2016] - Learning Deep Features for Discriminative Localization.
[Zhou et al. 2018] - Weakly Supervised Instance Segmentation using Class Peak Response
[Cholakkal et al. 2019] - Object Counting and Instance Segmentation with Image-level Supervision
[Tang et al. 2017] - Multiple Instance Detection Network with Online Instance Classifier Refinement
[Dai et al. 2015] - BoxSup: Exploiting Bounding Boxes to Supervise Convolutional Networks for Semantic Segmentation

issamou@cs.ubc.ca, dvazquez@elementai.com, schmidtm@cs.ubc.ca