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

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.

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

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.

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]

Experimental setup

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 a 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 mask of highest IoU

Quantitative results

Results on PASCAL VOC 2012 dataset

<table>
<thead>
<tr>
<th>Method</th>
<th>Supervision</th>
<th>mAP25</th>
<th>mAP50</th>
<th>mAP75</th>
<th>ABO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mask R-CNN</td>
<td>pixel-level</td>
<td>58.9</td>
<td>51.4</td>
<td>32.4</td>
<td>-</td>
</tr>
<tr>
<td>DeepMask</td>
<td>pixel-level</td>
<td>-</td>
<td>41.7</td>
<td>90.7</td>
<td>-</td>
</tr>
<tr>
<td>PRM</td>
<td>image-level</td>
<td>44.3</td>
<td>26.8</td>
<td>90.0</td>
<td>37.6</td>
</tr>
<tr>
<td>PRM+Density</td>
<td>image-level++</td>
<td>48.5</td>
<td>30.2</td>
<td>14.4</td>
<td>44.3</td>
</tr>
<tr>
<td>DeepMask</td>
<td>bounding box</td>
<td>39.4</td>
<td>0.81</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>WISE (Ours)</td>
<td>image-level</td>
<td>48.2</td>
<td>40.4</td>
<td>22.2</td>
<td>51.3</td>
</tr>
<tr>
<td>WISE+Refine</td>
<td>image-level</td>
<td>48.2</td>
<td>41.7</td>
<td>23.7</td>
<td>55.2</td>
</tr>
</tbody>
</table>

Qualitative results

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.

References

He et al. 2017: Mask R-CNN.
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: Bbox supervision for semantic segmentation.