Scaling up Instance Segmentation using *Approximately Localized Phrases*

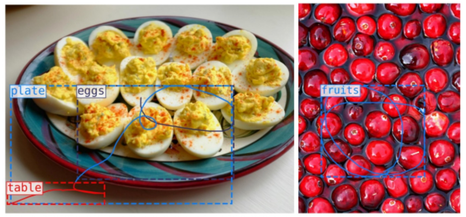
Karan Desai [1] Ishan Misra [2] Justin Johnson [1,2] Laurens van der Maaten [2]

[1] University of Michigan [2] Meta AI

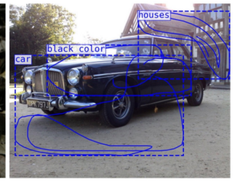
BMVC 2022

1 Training data with ALPs

ALPs: natural language phrases paired with boxes derived from mouse scribbles.



Source for ALPs:
Localized Narratives
(COCO and Open Images)

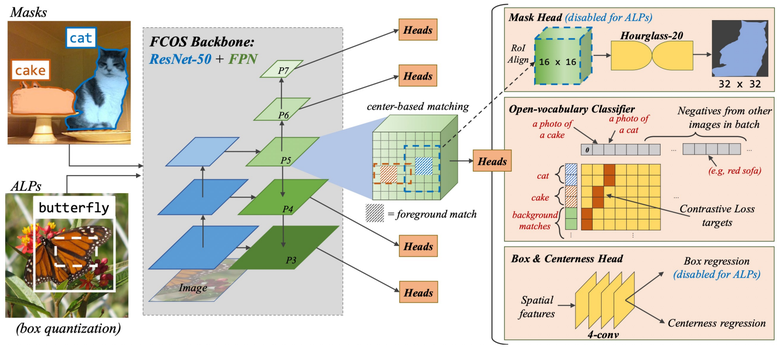


LocNar-OID stats:
- *504K images*
- *2.75M ALPs*
- *51K unique phrases*

240 novel classes that do not have training masks or boxes in COCO.

2 Model: FCOS-MO

FCOS-style model with a deep Hourglass mask head and a CLIP-based open-vocabulary classifier.



FCOS uses box centers for localization supervision during training instead of box area or edges as used by Mask R-CNN. This makes it less sensitive to noise in imprecise ALP boxes.

- ***Training:*** *alternate batches of masks (base classes) and ALPs (novel classes).*
- ***Box quantization:*** *Training augmentation to randomly expand boxes of ALPs. Increases the coverage of underlying object in the box, provides less noisy training supervision.*

Motivation and Goal

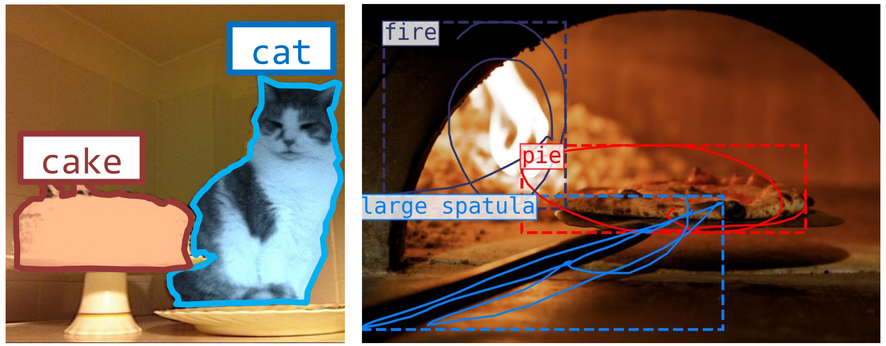
Labeled masks and boxes are expensive to collect.

*Learn to segment **novel object classes** without using any training masks or labeled boxes for them.*

Main Idea

Training instance segmentation models with:

Mask annotations *(base classes)* + *Approximately Localized Phrases (novel classes)*



3 Quantitative Results

Main experiments: *Train Mask R-CNN* (baseline) and FCOS-MO models using COCO masks and different forms of supervision for novel classes, for learning to segment 300 classes from Open Images* ***(60 base, 240 novel)****.*

Evaluation: *Open Images test AP_{50} for base and novel classes.*

Model	Novel class supervision	AP_{50}-base	AP_{50}-novel
Open-vocab Mask R-CNN*	none	62.8	21.4
ALP-supervised Mask R-CNN*	ALPs	56.7	27.8 **(+6.4)**
ALP-supervised FCOS-MO	ALPs	60.0	30.5 **(+9.1)**
Box-supervised Mask R-CNN* **(oracle)**	Labeled boxes	61.8	50.5

Observation 1: *ALP-supervised models outperform an open-vocab baseline.*
Observation 2: *When trained with ALPs, FCOS-MO outperforms Mask R-CNN* indicating its effectiveness in handling noisy supervision of ALPs.*

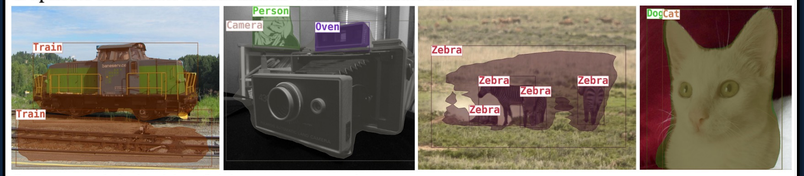
… see our paper for evaluations on COCO subsets and ablation studies.

4 Qualitative Results

Predictions on Open Images from ALP-supervised FCOS-MO.

Accurate labels with accurate masks:

Duplicate inaccurate labels/masks with an accurate mask:



Inaccurate label with a fairly accurate mask: