Learning to Detect Every Thing in an Open World

Kuniaki Saito[1] , Ping Hu[1], Trevor Darrell[3] and Kate Saenko[1,2]

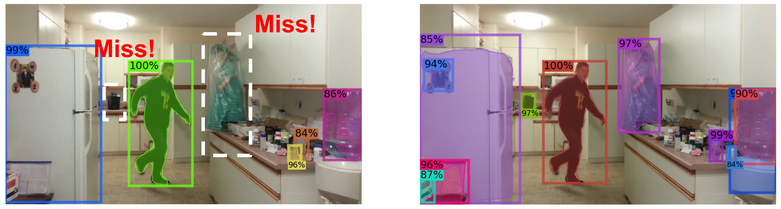
1: Boston University, 2: MIT-IBM Watson AI Lab, 3: University of California Berkeley

Introduction

Open-world Instance Segmentation

- Detect and segment novel objects as “object”
- Novel = categories not covered by dataset

Mask RCNN (trained on COCO) **Ours (trained on COCO)**



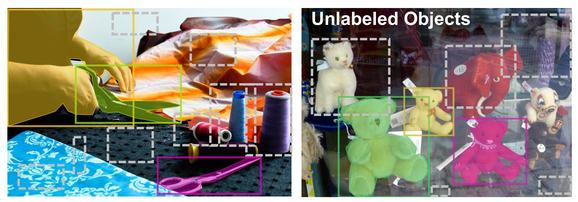
Issues

- Dataset does not cover all categories of objects

⇒ Many unlabeled objects in images

- Regard the unlabeled objects as background when training detector

⇒ Objectness of novel objects is suppressed.



Proposed Approach

- We do not want to suppress unlabeled objects as background.

⇒ ***Background Erasing: Erase unlabeled objects***

- Need to solve domain-gap between synthesized and real data

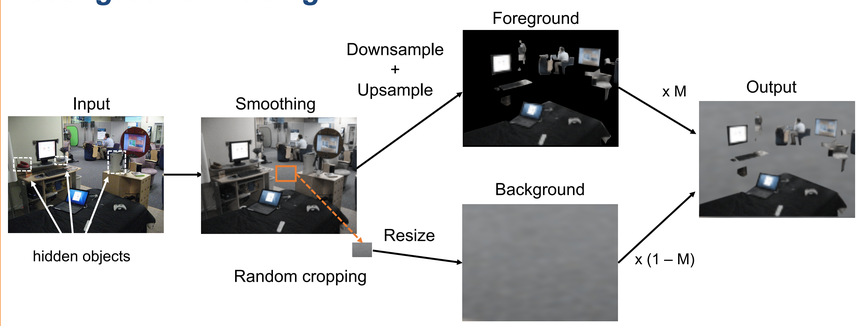
⇒ ***Decoupled Training: Solve the domain-gap***

Summary

- A simple framework for open-world instance segmentation.
- This method is effective in one-stage and two-stage detectors.
- Applicable to diverse datasets, e.g., COCO, UVO, Mapillary
- 12.8 points improvement on VOC to Non-VOC evaluation.

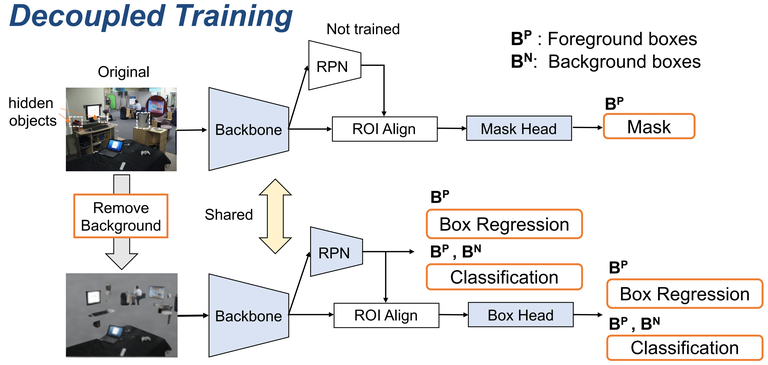
Method

Background Erasing



- Copy-paste foreground object on background canvas.

Decoupled Training



- Detection loss on synthesized images.
- Segmentation loss on real images.

Results

Cross-Category Evaluation on COCO

Method	Non-VOC Box AP	Non-VOC Box AR_{10}	Non-VOC Box AR_{100}	Non-VOC Mask AP	Non-VOC Mask AR_{10}	Non-VOC Mask AR_{100}	All Box AR_{10}	All Box AR_{100}	All Mask AR_{10}	All Mask AR_{100}
Mask RCNN	1.5	8.8	10.9	0.7	7.2	9.1	19.3	23.1	16.7	19.9
Mask RCNN^P	3.4	8.7	10.7	2.2	7.2	8.9	19.1	23.0	16.5	19.8
Mask RCNN^S	1.1	13.2	18.0	0.6	11.3	15.8	21.7	27.4	19.2	24.4
LDET	**5.0**	**18.2**	**30.8**	**4.7**	**16.3**	**27.4**	**24.4**	**36.8**	**22.4**	**33.1**

Ablation Study

Method	Box Real	Box Synth	Mask Real	Mask Synth	Box AR_{10}	Box AR_{100}	Box AR0.5
Plain	✓		✓		8.8	10.9	19.1
Synth Only		✓			1.6	4.3	11.7
Synth Only*		✓		✓	3.0	9.5	23.8
LDET		✓	✓		**18.2**	**30.8**	**53.2**

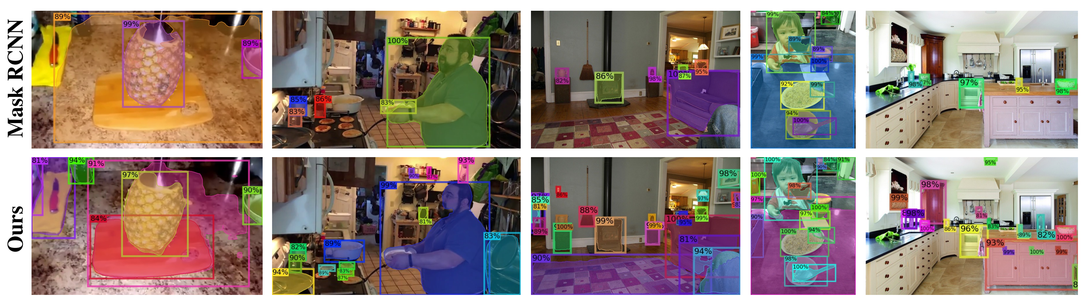
Cross-Dataset Evaluation on COCO to UVO

Method	Train	Box AP	Box AR	Box AR_{small}	Box AR_{med}	Box AR_{large}	Mask AP	Mask AR	Mask AR_{small}	Mask AR_{med}	Mask AR_{large}
Mask RCNN	VOC (COCO)	19.8	30.0	10.7	21.3	43.0	15.5	23.9	9.2	18.5	32.8
Mask RCNN^P		19.2	30.1	10.6	21.3	43.3	15.4	24.1	9.4	18.4	33.2
Mask RCNN^S		19.7	32.0	10.0	23.3	46.0	14.1	25.9	9.5	20.2	35.4
LDET		**22.4**	**43.7**	**24.7**	**39.9**	**52.9**	**18.4**	**36.0**	**22.1**	**34.8**	**41.4**
Mask RCNN	COCO	25.3	42.3	22.2	38.3	52.0	20.6	35.9	19.6	33.9	42.6
Mask RCNN^P		24.4	41.9	22.3	37.8	51.5	20.1	35.4	19.7	33.6	41.8
Mask RCNN^S		23.4	40.5	17.6	34.9	52.3	18.0	34.7	16.6	31.5	42.8
LDET		**25.8**	**47.5**	**29.1**	**44.8**	**55.6**	**21.9**	**40.7**	**26.8**	**40.0**	**45.7**

Evaluation on One-stage detector

Detector	Method	AR_{10}	AR_{50}	AR_{100}
RetinaNet	Plain	9.9	15.7	17.8
	LDET	**15.3**	**26.7**	**31.0**
TensorMask	Plain	10.6	17.6	19.7
	LDET	**16.3**	**26.8**	**31.1**

COCO to UVO



Objectness