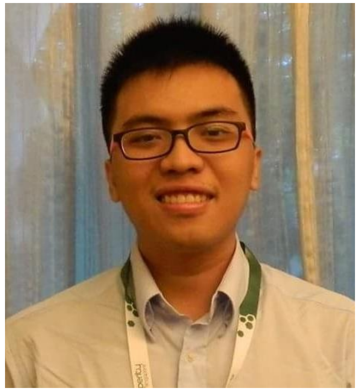
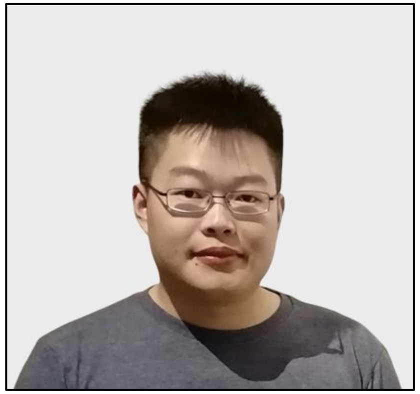
Open-Vocabulary Instance Segmentation via Robust Cross-Modal Pseudo-Labeling



Dat Huynh [1]

Jason Kuen [2]

Zhe Lin [2]



Jiuxiang Gu [2]

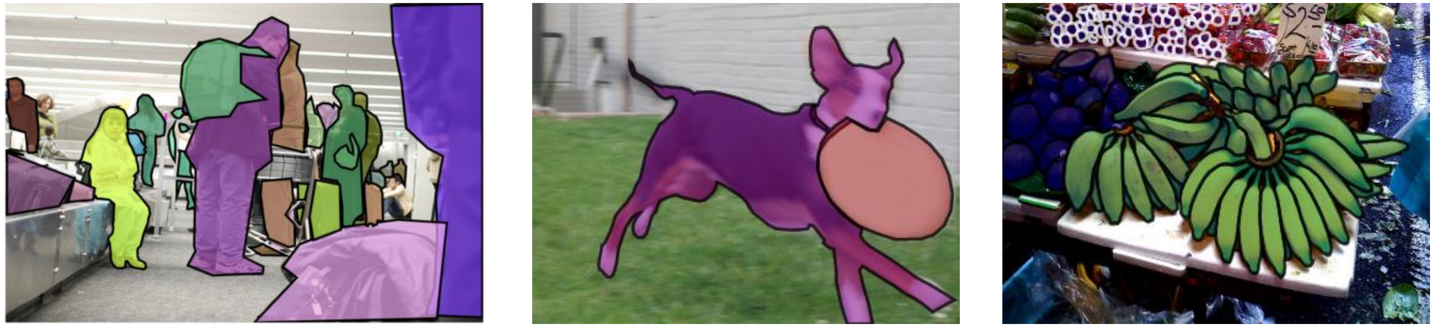
Ehsan Elhamifar [1]

[1] Northeastern University

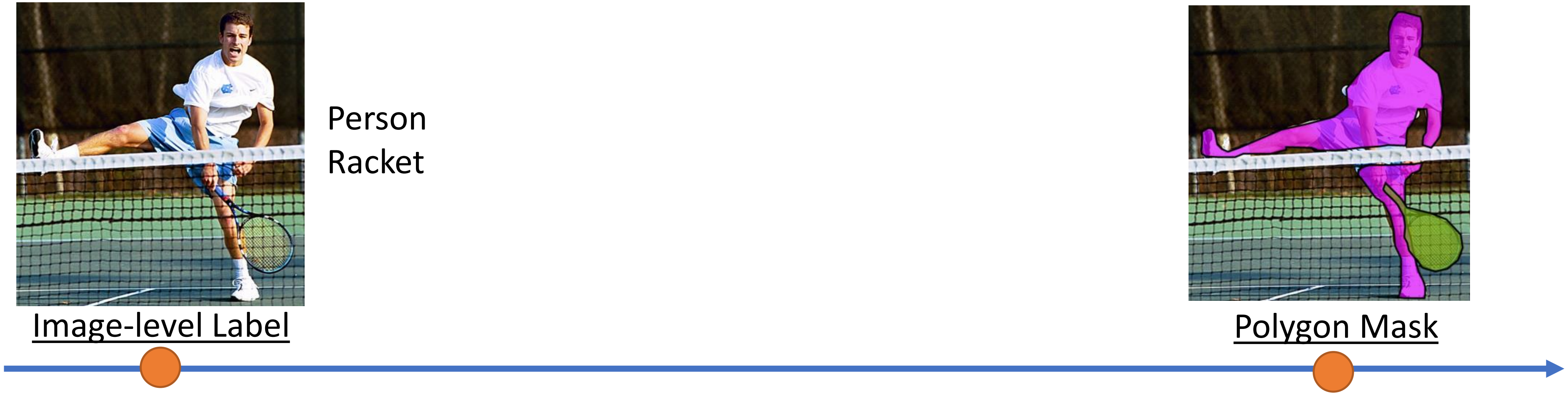
[2] Adobe Research

Instance Segmentation

- Instance Segmentation: segment every object in image



- *Costly* to annotate masks for many classes



Person
Racket

Image-level Label

Polygon Mask

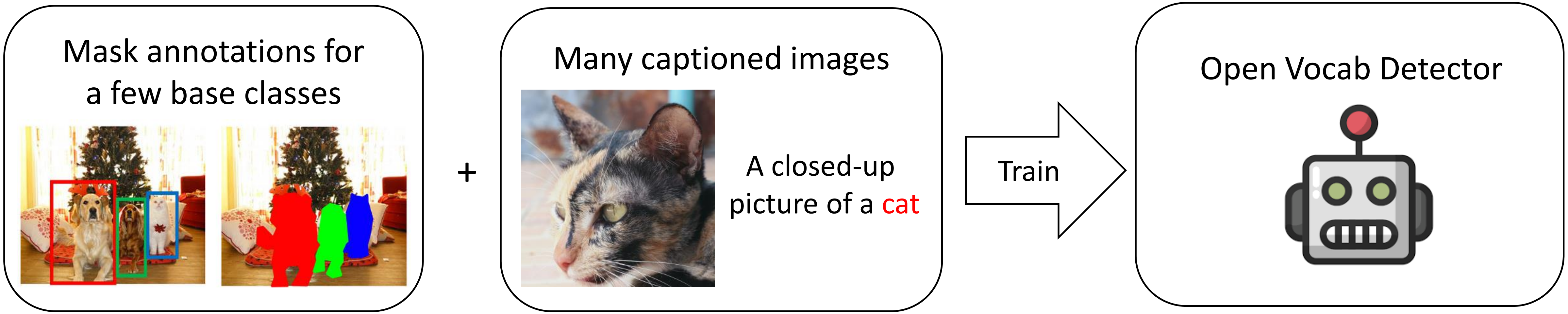
1s per label

80s per instance

Lin et al., ECCV 14

Mixed Supervision

- Use ***mixed supervision*** for training



Mask annotations for a few base classes

+

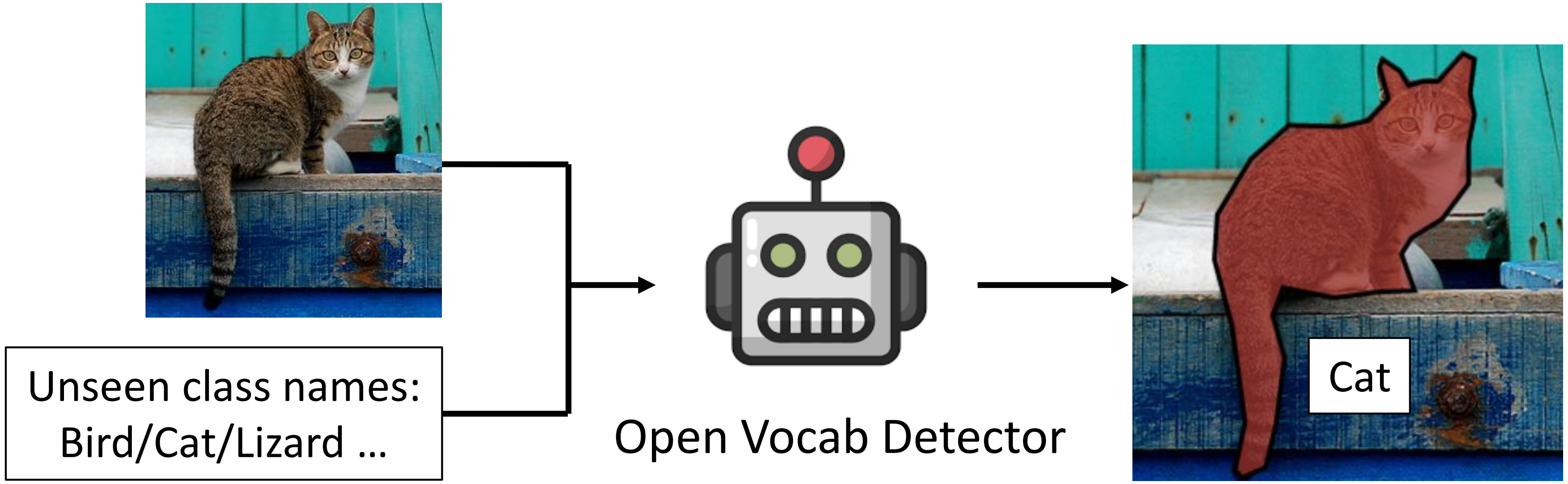
Many captioned images

A closed-up picture of a cat

Train

Open Vocab Detector

- Segment classes without any mask annotations



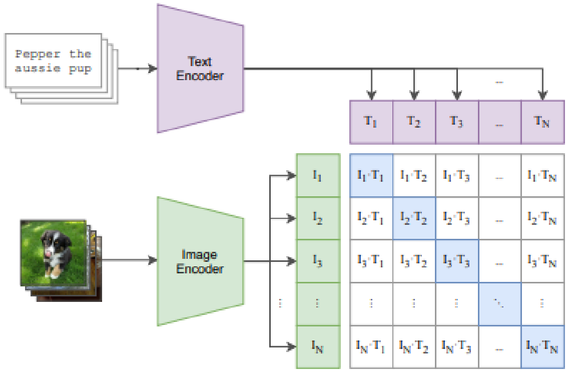
Unseen class names: Bird/Cat/Lizard …

Open Vocab Detector

Cat

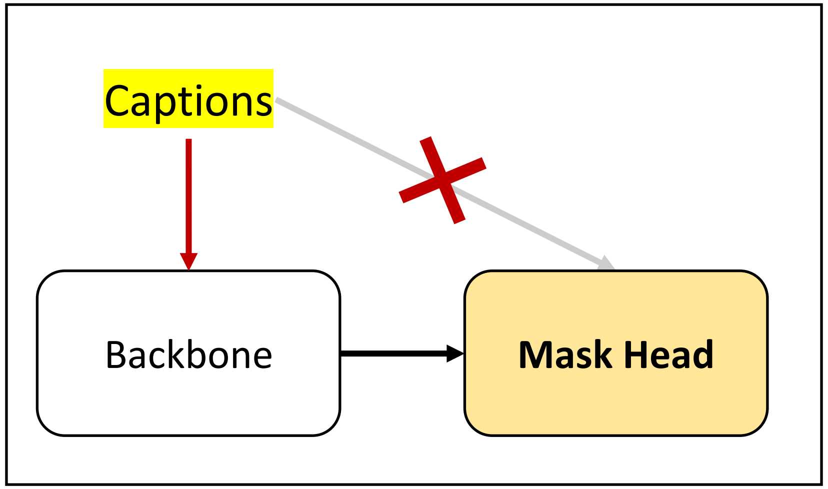
Prior Works

- Contrastive Representation Learning
  - Maximize similarity between corresponding {image, caption}


Pepper the aussie pup
Text Encoder
Image Encoder
T1 T2 T3 … TN
I1 I2 I3 ⋮ IN
I1·T1 I1·T2 I1·T3 … I1·TN
I2·T1 I2·T2 I2·T3 … I2·TN
I3·T1 I3·T2 I3·T3 … I3·TN
IN·T1 IN·T2 IN·T3 … IN·TN

Radford et. al. ICML21

- Prior Work: Backbone Pretraining
  - ***Cannot use captions*** to train mask head

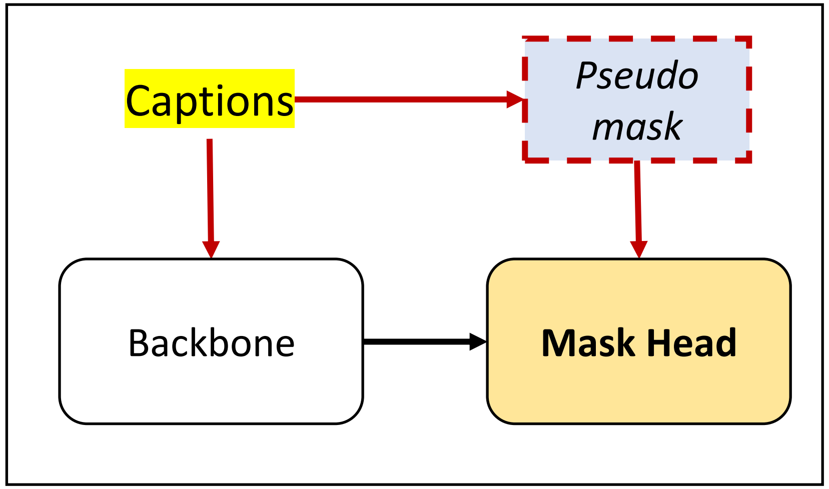


Captions

Backbone

Mask Head

- Ours: ***Pseudo masks***



Captions

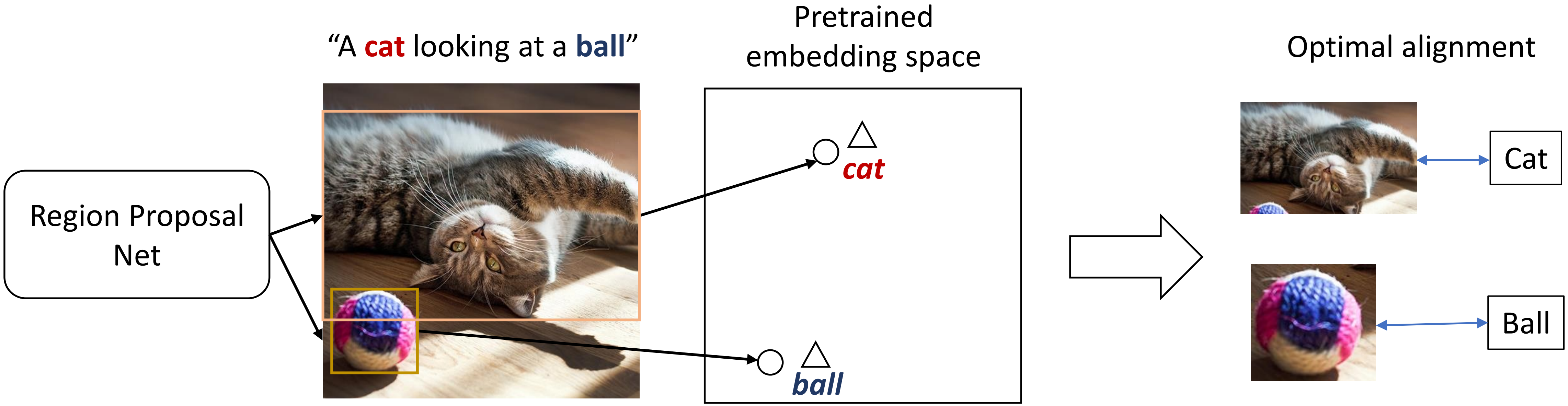
Pseudo mask

Backbone

Mask Head

Zero-shot Teacher

- Cross-Modal Alignment:



"A **cat** looking at a **ball**"

Region Proposal Net

Pretrained embedding space

cat

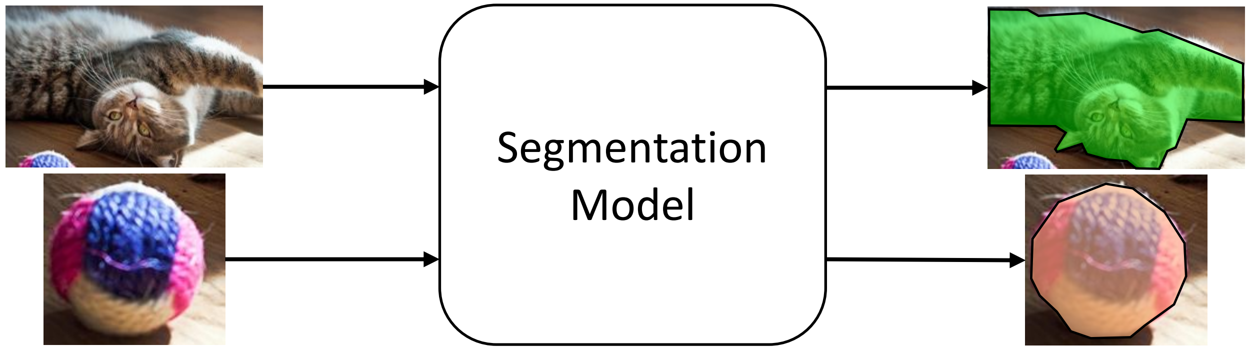
ball

Optimal alignment

Cat

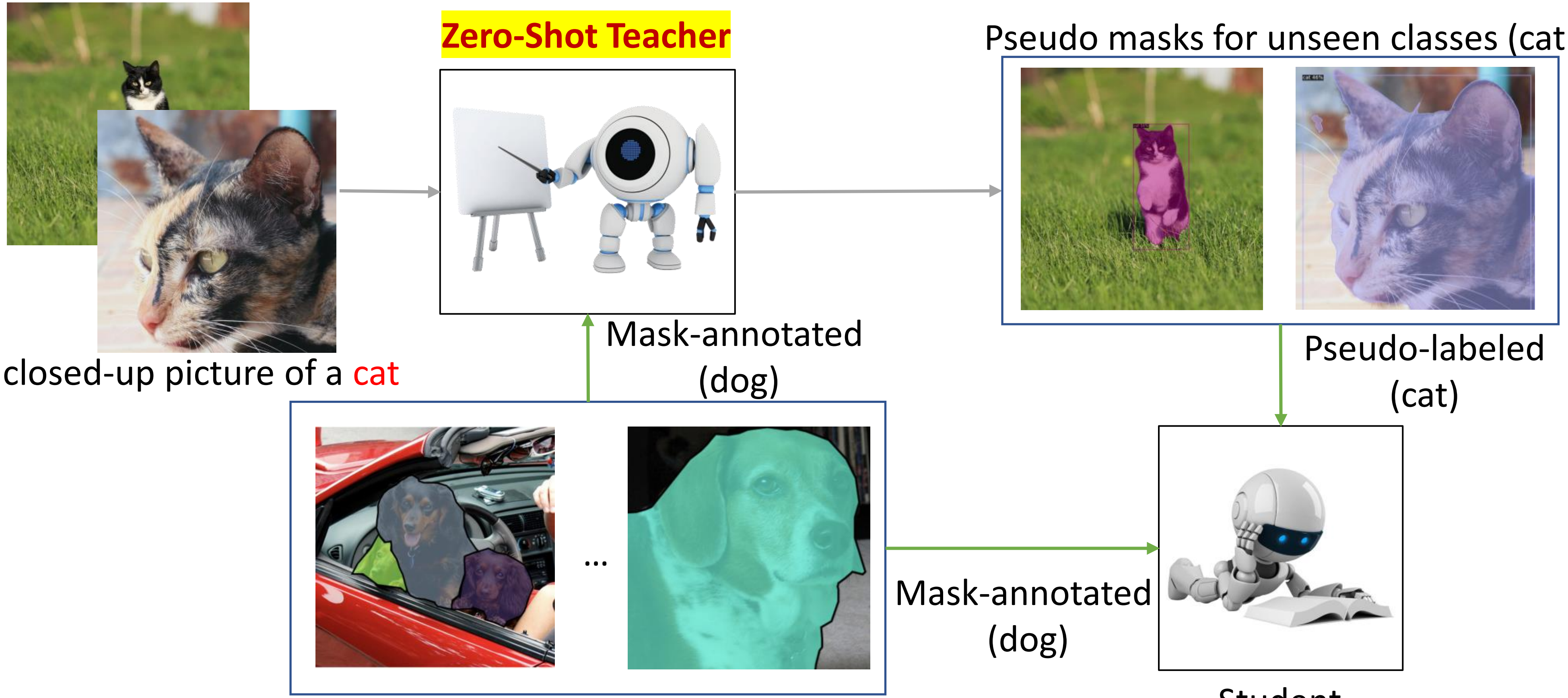
Ball

- Class-Agnostic Segmentation:



Segmentation Model

Student-Teacher Framework



A black and white cat running on grass

A closed-up picture of a cat

Zero-Shot Teacher

Pseudo masks for unseen classes (cat)

Mask-annotated (dog)

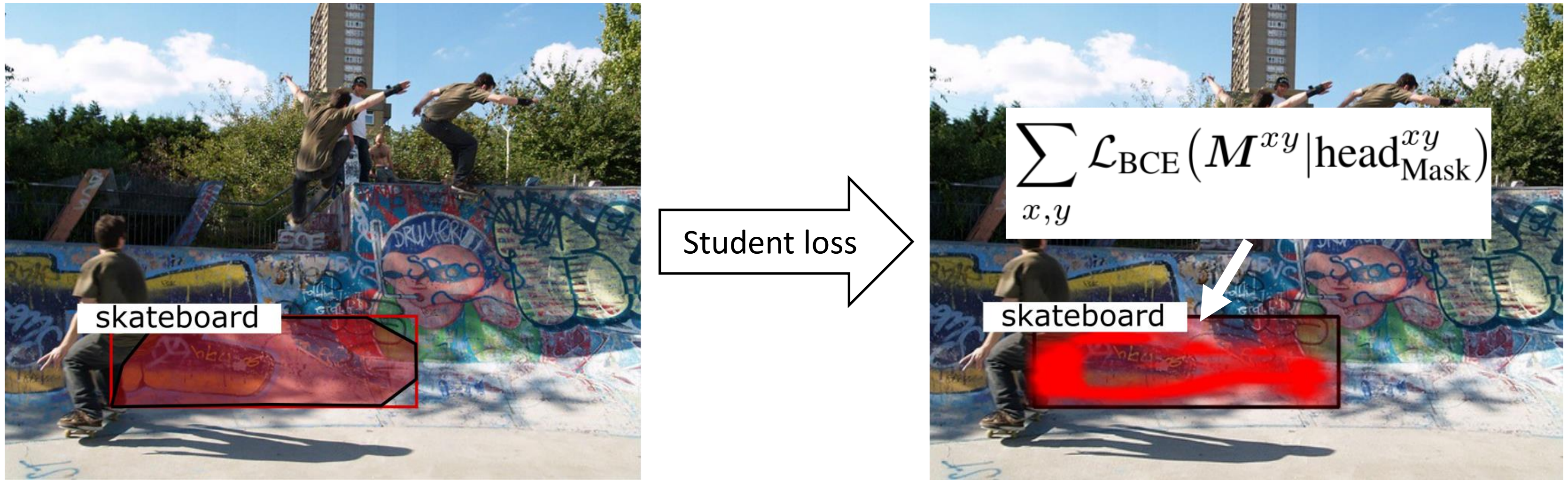
…

Pseudo-labeled (cat)

Mask-annotated (dog)

Student

Pseudo-Label Errors



Student loss

$$\sum_{x,y} \mathcal{L}_{\text{BCE}}\left(\boldsymbol{M}^{xy} | \text{head}^{xy}_{\text{Mask}}\right)$$

- Mask noise estimation:

$$\sum_{x,y} \mathcal{L}_{\text{BCE}}\left(\boldsymbol{M}^{xy} | \text{head}^{xy}_{\text{Mask}} \boxed{+ \epsilon^{xy}}\right)$$

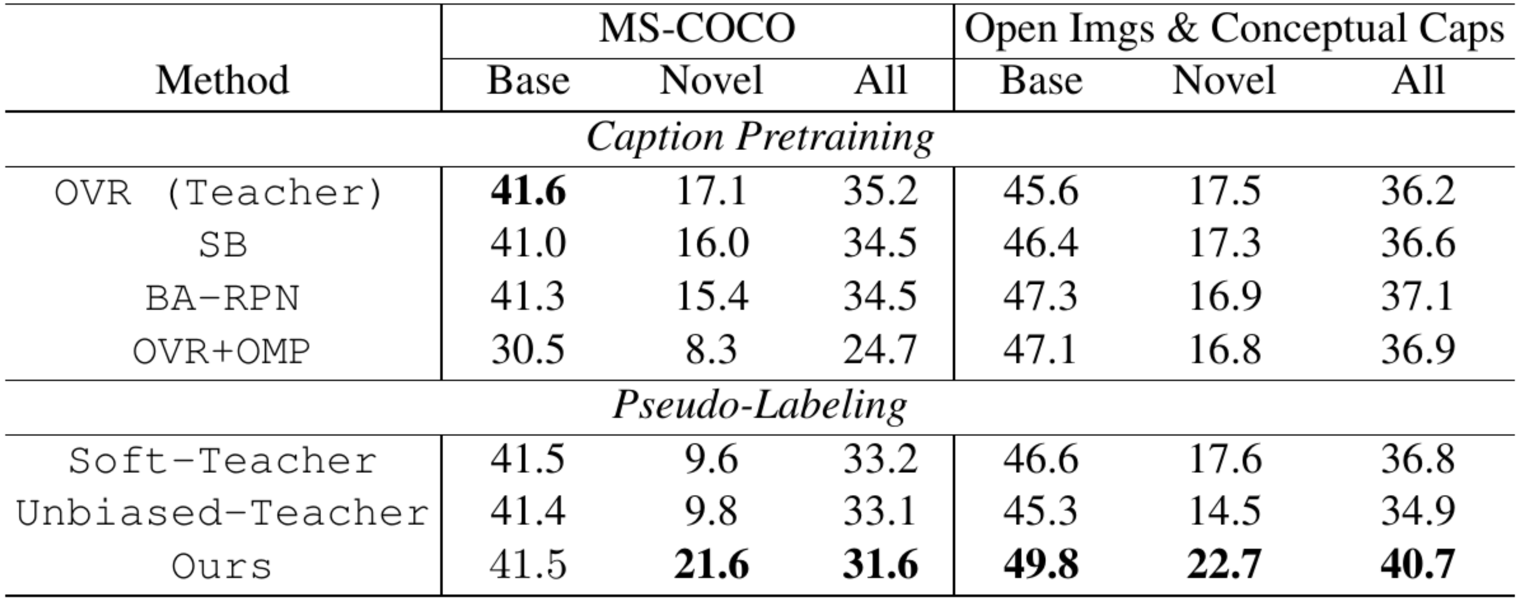
$$\epsilon^{xy} \sim \mathcal{N}\left(0, \text{head}^{xy}_{\text{Noise}}\right)$$

- Loss reweighting:

$$\alpha(\boldsymbol{M}) \propto \frac{1}{\sum_{x,y} \text{head}^{xy}_{\text{Noise}} / |\boldsymbol{M}|}$$

Experimental Results

- Main experiments
  - Metric: mAP
  - MS-COCO: 48 base/ 17 novel
  - Open Images: 200 base/ 100 novel



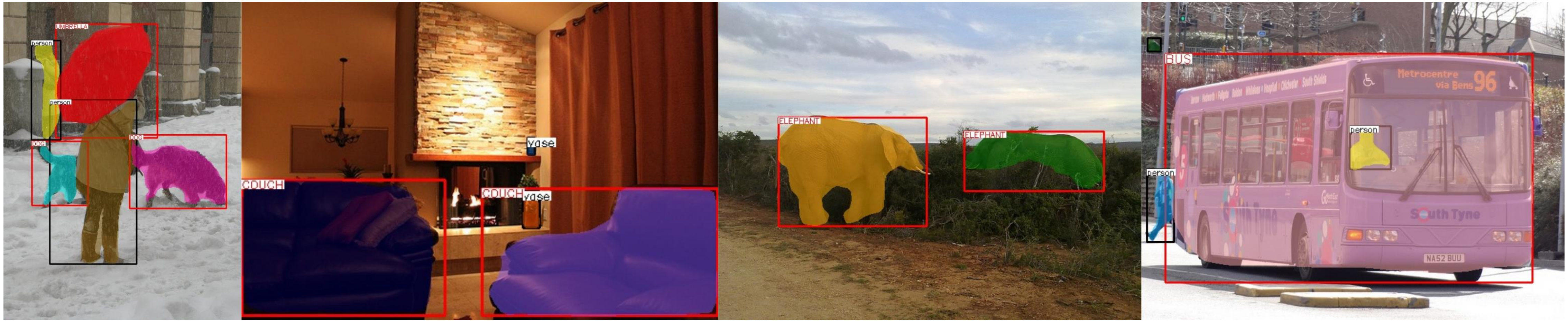
Method	MS-COCO			Open Imgs & Conceptual Caps		
	Base	Novel	All	Base	Novel	All
Caption Pretraining						
OVR (Teacher)	**41.6**	17.1	35.2	45.6	17.5	36.2
SB	41.0	16.0	34.5	46.4	17.3	36.6
BA-RPN	41.3	15.4	34.5	47.3	16.9	37.1
OVR+OMP	30.5	8.3	24.7	47.1	16.8	36.9
Pseudo-Labeling						
Soft-Teacher	41.5	9.6	33.2	46.6	17.6	36.8
Unbiased-Teacher	41.4	9.8	33.1	45.3	14.5	34.9
Ours	41.5	**21.6**	**31.6**	**49.8**	**22.7**	**40.7**

+4.5%

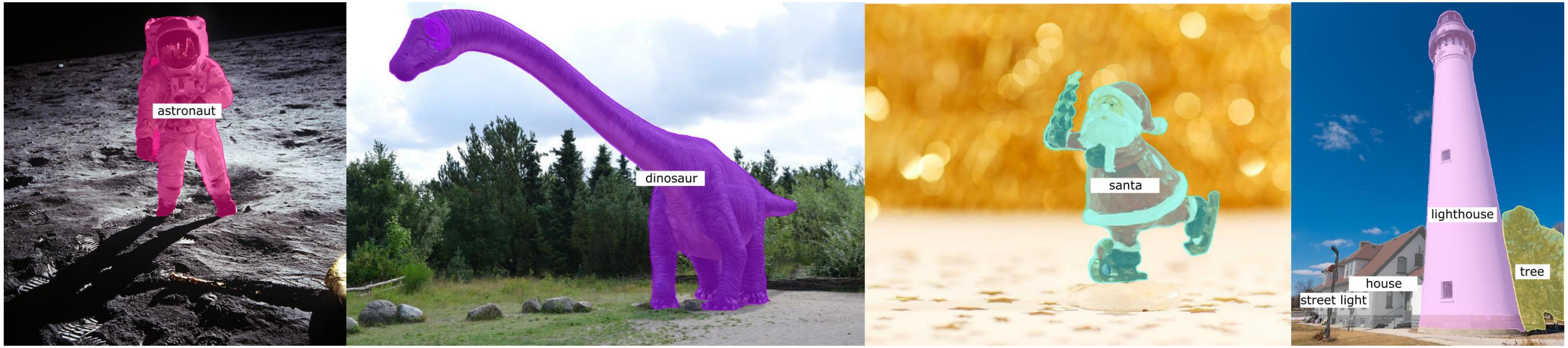
+5.1%

Qualitative Results

- MS-COCO:



- Unseen object in the wild:



Code

Code is available at:

https://github.com/hbdat/cvpr22_cross_modal_pseudo_labeling