Scaling up Instance Segmentation using Approximately Localized Phrases

Supplementary material

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

[1] University of Michigan
[2] Meta AI

The supplementary material is organized as follows:
- In Sec. 1, we include extra training details and evaluation protocol of main experiments.
- In Sec. 2, we show randomly selected qualitative outputs from our model.
- In Sec. 3, we provide a plausible explanation to why FCOS-MO is less sensitive to noisy boxes, by simulating imprecise boxes in COCO.

1 Training Setup: Additional Details

Our training setup involves learning to segment $|\mathcal{C}| = 300$ classes from Open Images. We reported two separate metrics, AP_{50}-*base* and AP_{50}-*novel* for $|\mathcal{B}| = 60$ *base* and $|\mathcal{N}| = 240$ *novel* classes respectively. The tables below list all these classes in alphabetical order. Class names mentioned with a blue italic style are *novel*, remaining are *base* classes.

Out of 60 *base* classes, 51 classes occur as-is in COCO class ontology. We included five *base* classes that are differently worded but semantically identical: *Doughtnut* ⇒ *donut*, *Computer mouse* ⇒ *mouse*, *Computer keyboard* ⇒ *keyboard*, *Microwave oven* ⇒ *microwave*, and *Kitchen knife* ⇒ *knife*. Finally, we also marked four OID classes *Man, Woman, Boy, Girl* as *base* classes, since the *person* class in COCO provides abundant mask annotations.

Adhesive tape	*Aircraft*	Airplane	*Alarm clock*
Alpaca	*Ambulance*	Apple	*Asparagus*
Backpack	*Bagel*	*Ball*	*Balloon*
Banana	*Barge*	*Barrel*	Baseball bat
Baseball glove	*Bat*	*Beaker*	Bear
Beer	*Bell pepper*	*Belt*	*Bicycle wheel*
Billiard table	*Binoculars*	Bird	*Blue jay*
Book	*Boot*	Bottle	Bowl
Box	Boy	*Bread*	*Briefcase*
Broccoli	*Bronze sculpture*	*Brown bear*	*Bull*

Table 1: **OID classes, part 1.** *Blue italics* indicate *novel* classes (240 in total).

Burrito	Bus	*Bust*	*Cabbage*
Cake	*Camel*	*Camera*	*Canary*
Candle	*Canoe*	Car	*Carnivore*
Carrot	Cat	*Cattle*	*Cello*
Cheetah	*Chest of drawers*	*Chicken*	*Chopsticks*
Christmas tree	Clock	*Cocktail*	*Coffee*
Coffee cup	*Coin*	*Common fig*	Computer keyboard
Computer mouse	*Cookie*	*Corded phone*	Couch
Cowboy hat	*Crocodile*	*Croissant*	*Cucumber*
Dagger	*Dice*	*Digital clock*	Dog
Dog bed	*Dolphin*	*Door handle*	Doughnut
Dress	*Drink*	*Drinking straw*	*Duck*
Dumbbell	*Eagle*	Elephant	*Envelope*
Falcon	*Fedora*	*Filing cabinet*	Fire hydrant
Fish	*Flag*	*Flashlight*	*Flower*
Flowerpot	*Flute*	*Food processor*	*Football*
Fox	*Frog*	*Frying pan*	Giraffe
Girl	*Glove*	*Goat*	*Goldfish*
Goose	*Grape*	*Grapefruit*	*Guacamole*
Guitar	*Hamburger*	*Hamster*	Handbag
Handgun	*Harbor seal*	*Harpsichord*	*Hat*
High heels	Horse	Hot dog	*Human ear*
Human mouth	*Jaguar*	*Jeans*	*Jet ski*
Jug	*Juice*	*Kangaroo*	*Kettle*
Kitchen knife	Kite	Knife	Laptop
Lemon	*Leopard*	*Light switch*	*Lighthouse*
Limousine	*Lion*	*Lizard*	*Loveseat*
Luggage and bags	*Lynx*	Man	*Mango*
Microwave oven	*Miniskirt*	*Missile*	*Mobile phone*
Monkey	Motorcycle	*Mouse*	*Muffin*
Mug	*Mule*	*Mushroom*	*Nail*
Orange	*Ostrich*	*Otter*	Oven
Owl	*Oyster*	*Pancake*	*Paper towel*
Parrot	*Peach*	*Pear*	*Pen*
Penguin	Person	*Piano*	*Picture frame*
Pig	*Pillow*	*Pitcher*	Pizza
Plastic bag	*Platter*	*Polar bear*	*Pomegranate*
Potato	*Power plugs and sockets*	*Pressure cooker*	*Pretzel*
Printer	*Pumpkin*	*Punching bag*	*Rabbit*
Raccoon	*Racket*	*Radish*	*Raven*
Reptile	*Rhinoceros*	*Rocket*	*Roller skates*
Rose	*Rugby ball*	*Ruler*	Sandwich
Saucer	*Saxophone*	*Scarf*	Scissors
Screwdriver	*Sculpture*	*Sea lion*	*Sea turtle*
Seat belt	*Segway*	*Shark*	Sheep
Shirt	*Shorts*	*Shower*	Skateboard
Skirt	*Skull*	*Skyscraper*	*Slow cooker*
Snake	*Snowmobile*	*Sock*	*Sofa bed*
Sombrero	*Sparrow*	*Spatula*	Spoon
Squash	*Squirrel*	*Starfish*	Stop sign
Strawberry	*Studio couch*	*Submarine sandwich*	*Suit*
Suitcase	*Sun hat*	*Sunflower*	Surfboard
Swan	*Swim cap*	*Swimwear*	*Sword*
Table tennis racket	*Tablet computer*	*Tank*	*Tap*
Tart	*Taxi*	*Tea*	*Teapot*
Teddy bear	*Tennis ball*	Tennis racket	Tie
Tiger	Toaster	Toilet	*Toilet paper*

Table 2: **OID classes, part 2.** *Blue italics* indicate *novel* classes (240 in total).

Tomato	*Torch*	*Tortoise*	*Towel*
Toy	Traffic light	*Traffic sign*	Train
Trousers	Truck	*Turkey*	*Turtle*
Van	Vase	*Vehicle registration plate*	*Volleyball*
Waffle	*Washing machine*	*Waste container*	*Watch*
Watermelon	*Whale*	*Wheel*	*Whiteboard*
Wine	*Winter melon*	*Wok*	Woman
Woodpecker	*Wrench*	Zebra	*Zucchini*

Table 3: **OID classes, part 3.** *Blue italics* indicate *novel* classes (240 in total).

Extracting phrases from LocNar captions: Sec 3.1 of the main paper explains how we obtained ALPs from LocNar-OID. Phrases in ALPs are chunks of consecutive *adjectives* and *nouns* occurring in captions. LocNar captions often mention words like *sky* and *water*, which are commonly called *stuff* classes, and not targetted for instance segmentation. Captions also contain words that convey position, orientation, and spatial arrangement of objects, for example "In the *background* I see buildings...". Such abstract nouns do not provide any meaningful supervision for segmenting individual *thing* objects.

We manually define a blocklist of noun words and discard phrases that contain them. Our filtering is very lightweight as compared to deciding full-scale label ontologies. Note that LocNar was not collected with an intent to target instance-segmentation – such filtering can be minimized by incorporating well-defined instructions during data collection.

Stuff classes and very generic words				
air	ceiling	cloud	floor	grass
ground	group	ice	item	lake
land	object	ocean	platform	river
road	room	sand	sea	sky
snow	stone	surface	wall	water
Nouns conveying spatial relationships				
foreground	background	top	bottom	left
right	middle	center	front	back
side	backside	inside	outside	outdoor
indoor	corner	area	view	
Meta information about the image file				
image	blur	watermark	text	
Reference to time and weather				
day	night	time	daytime	nighttime

Table 4: **Blocklist of nouns** while extracting phrases from LocNar-OID.

2 Qualitative Examples

In the main paper, we discussed common success and failure modes of our ALP-supervised FCOS-MO with selected qualitative predictions. Here we provide some *randomly selected* qualitative examples in Fig. 1. All masks were predicted with at least 20% confidence threshold and filtered by applying NMS with 0.2 IoU threshold. These results suggest that object classification is a more challenging subtask than mask prediction – our model tends to predict high-quality masks despite failing to predict accurate labels (last row). We view this as a positive indication, that instance segmentation models can be scaled beyond existing datasets like COCO simply by scaling up language supervision rather than using more masks.

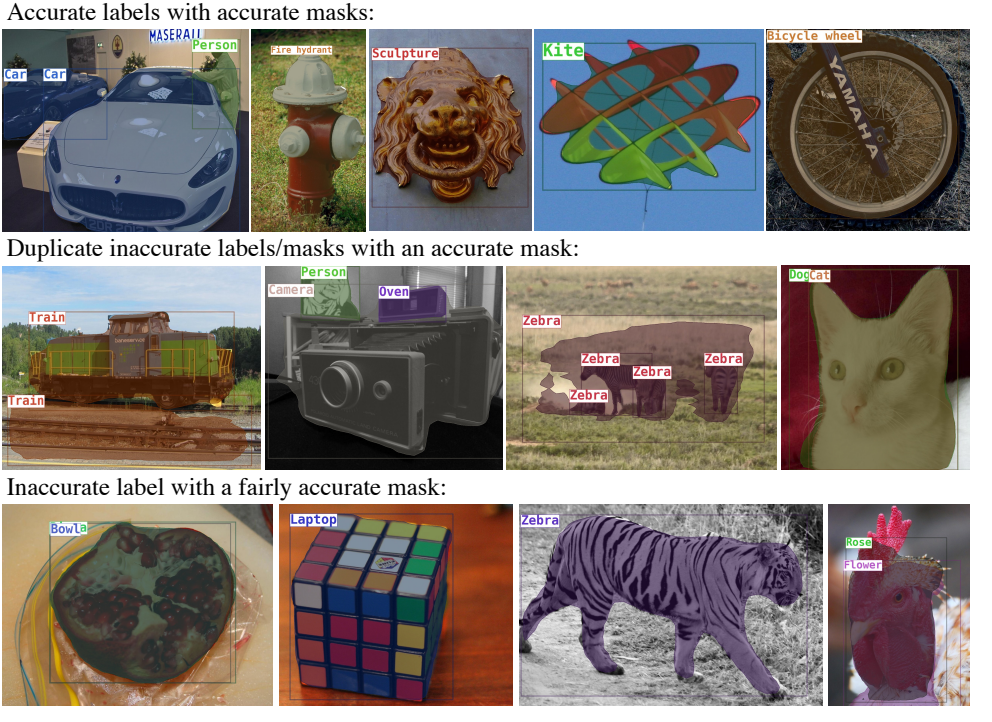
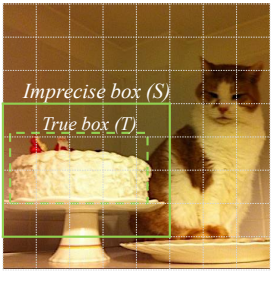


Figure 1: **FCOS-MO predictions:** Our model predicts quite accurate masks for diverse visual objects. Incorrect predictions (second and third rows) largely suggest that our model fails at classifying more often than failing to predict high-quality masks.

Per-class pooled predictions: We show more qualitative mask predictions for novel classes by pooling them across validation set – for each object class, we obtain predicted masks on OID-v6 val split, sort them by confidence score, and randomly select 4–5 predictions out of top-10. Fig. 2 below shows mask predictions for a few more classes, similar to those in main paper. We find that pooled mask predictions, even though selected randomly, are quite accurate – indicating that the model predicts accurate masks and labels with high confidence. However our model is not perfect, we observe similar failure modes as earlier, where either the predicted class or the quality of predict mask is inaccurate.

Figure 2: **FCOS-MO outputs obtained by pooling per-class predictions.**

3 Understanding FCOS-MO's Robustness to Noisy Boxes



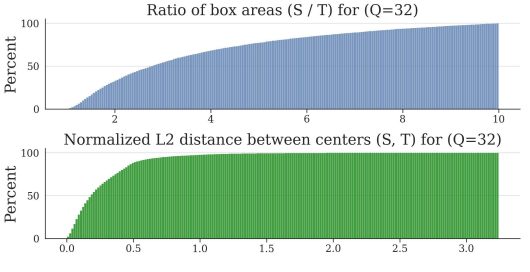


Figure 3: **Understanding FCOS-MO's robustness to noisy boxes:** We observe the cumulative distributions of area ratio (*blue*) and L2 distances between centers (*green*) of simulated and true boxes in COCO. The shift in box centers is less than distortion in area of boxes.

One major difference in modeling design of FCOS as compared to R-CNN style detectors is the use of localization supervision based on bounding box centers. FCOS performs *center-based matching* for supervision, and also adds an auxiliary centerness objective that uses box centers to learn localization.

We simulate imprecise boxes with COCO dataset. As a simple way to distort box edges, we perform box quantization as used in training FCOS-MO with a fixed $Q = 32$. Consider the example in Figure 3 (left) with an imprecise box (*cake*). Let its area be A_S and the area of

the true box in COCO be A_T. Figure 3 (right,top) shows the cumulative distribution of ratio (A_S/A_T) for all boxes in COCO. We observe that $> 50\%$ boxes are $> 3\times$ larger than original boxes. IoU-based matching is sensitive to GT box area – such imprecise boxes may lead to incorrect anchor assignments for RPN, and further have cascaded effect in the second stage.

On the other hand, FCOS matches features based on the their proximity to GT box centers [1, 2]. Figure 3 (right,bottom) shows the cumulative distribution of L2 distance $d(C_S, C_T)$ between centers of original and simulated boxes, normalized by size of the original box. This distribution is much steeper than area ratio, which suggests that FCOS may possibly be well suited for dealing with boxes with imprecise edges.

References

[1] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. FCOS: A Simple and Strong Anchor-Free Object Detector. *TPAMI*, 2020. 6

[2] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019. 6