Zero-Shot Instance Segmentation APPENDIX

Algorithm 1 Datasets construction process.

Input: The training set $COCO_{train}$ and validation set $COCO_{val}$ for MS-COCO 2014, the seen classes C_s and unseen classes C_u;

1: Select all images I_s from $COCO_{train}$ which contain objects belong to C_s;
2: Remove all the images which have any C_u object for I_s to construct training set D_{train};
3: Select all images I_u from $COCO_{val}$ which contain C_u classes objects to construct testing set D_{test};
4: **return** D_{train} and D_{test};

1. Dataset Description

Experiments are performed on two kind of splits for MS-COCO 2014: 48/17 split and 65/15 split, which mean 48 seen classes with 17 unseen classes and 65 seen classes with 15 unseen classes. The datasets construction process is shown in Alg 1. For training set, the 48/17 split has 44912 images and the 65/15 split has 82136 images. For the testing set, the 48/17 split has 2729 images and the 65/15 split has 10098 images.

For 48/17 split, the **seen** classes include: "person", "bicycle", "car", "motorcycle", "truck", "boat", "bench", "bird", "horse", "sheep", "zebra", "giraffe", "backpack", "handbag", "skis", "kite", "surfboard", "bottle", "spoon", "bowl", "banana", "apple", "orange", "broccoli", "carrot", "pizza", "donut", "chair", "bed", "tv", "laptop", "remote", "microwave", "oven", "refrigerator", "book", "clock", "vase", "toothbrush", "train", "bear", "suitcase", "frisbee", "fork", "sandwich", "toilet", "mouse", "toaster" and the **unseen** classes include: 'bus', 'dog', 'cow', 'elephant', 'umbrella', 'tie', 'skateboard', 'cup', 'knife', 'cake', 'couch', 'keyboard', 'sink', 'scissors', 'airplane', 'cat', 'snowboard'.

For 65/15 split, the **seen** classes include: 'person', 'bicycle', 'car', 'motorcycle', 'bus', 'truck', 'boat', 'traffic light', 'fire hydrant', 'stop sign', 'bench', 'bird', 'dog', 'horse', 'sheep', 'cow', 'elephant', 'zebra', 'giraffe', 'backpack', 'umbrella', 'handbag', 'tie', 'skis', 'sports ball', 'kite', 'baseball bat', 'baseball glove', 'skateboard', 'surfboard', 'tennis racket', 'bottle', 'wine glass', 'cup', 'knife', 'spoon', 'bowl', 'banana', 'apple', 'orange', 'broccoli', 'carrot', 'pizza', 'donut', 'cake', 'chair', 'couch', 'potted plant', 'bed', 'dining table', 'tv', 'laptop', 'remote', 'keyboard', 'cell phone', 'microwave', 'oven', 'sink', 'refrigerator', 'book', 'clock', 'vase', 'scissors', 'teddy bear', 'toothbrush' and the **unseen** classes include: 'airplane', 'train', 'parking meter', 'cat', 'bear', 'suitcase', 'frisbee', 'snowboard', 'fork', 'sandwich', 'hot dog', 'toilet', 'mouse', 'toaster', 'hair drier'.

Table 1. The results of the ablation experiments to the weight of reconstruct loss function in zero-shot detector and SMH. The Recall@100 results for ZSI and ZSD are reported on 48/17 split and 65/15 split of MS-COCO, respectively. λ_{ZSD} and λ_{SMH} are the weights for the reconstruct loss in zero-shot detector and SMH.

	λ_{ZSD}	λ_{SMH}	ZSI			ZSD
			0.4	0.5	0.6	0.5
48/17	0.5	0.5	50.3	44.9	38.7	53.9
	0.5	0.25	43.5	33.5	20.3	54.3
	0.5	1.0	51.6	45.4	39.0	54.8
	0.25	0.5	50.6	45.4	38.1	55.0
	1.0	0.5	51.1	44.6	37.6	54.9
65/15	0.5	0.5	55.8	50.0	42.9	58.9
	0.5	0.25	56.0	50.2	43.2	59.5
	0.5	1.0	57.1	51.8	45.1	58.6
	0.25	0.5	55.3	49.6	42.8	58.4
	1.0	0.5	57.1	51.4	44.4	60.1

2. Weight of the Reconstruction Loss Function

We report the influence of the weight of our reconstruction loss function on Table 1. We compare different weights for reconstruct loss function in zero-shot detector and SMH. We can learn that higher weight for reconstruction loss function can bring more performance improvement for detector and mask head.

3. Training hyperparameters

We train our zero-shot instance segmentation network using SGD with improved weight decay handling, set to

10^{-5} and the learning rate is 0.01. We also apply gradient clipping, with a maximal gradient norm of 35. The training schedule is a 12 epoch training process and reducing learning rate in 8 and 11 epoch by 10 factor. ImageNet pre-trained backbone ResNet-101 is imported from Torchvision, discarding the last classification layer. The batch normalization weights and statistics of backbone are frozen during training, following widely adopted practice in object detection. The first block for backbone are also frozen during training. We observe that clipping the gradient is important to stabilize training, especially in the first few epochs. The weights for the supplementary parameters are randomly initialized with normal initialization. We resize the shortest side for input images to 800 pixels with the longest at most 1333. A train image is horizontal flipped with probability 0.5 during training and we do not use any data augmentation in testing process.

4. Additional results

Some extra qualitative results for the prediction of the ZSI from our method are shown in Fig 1 and Fig 2. We can learn that our method can predict satisfactory results for both seen and unseen instances. And the failure cases mainly distributed in the misclassification between semantically similar categories.

Figure 1. Some extra qualitative results for the prediction of the ZSI.

Figure 2. Some extra qualitative results for the prediction of the ZSI.