Algorithm 1 Datasets construction process.

**Input:** The training set \textit{COCO}$_{\text{train}}$ and validation set \textit{COCO}$_{\text{val}}$ for MS-COCO 2014, the seen classes \( C_s \) and unseen classes \( C_u \);

1: Select all images \( I_s \) from \textit{COCO}$_{\text{train}}$ 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 \( D_{\text{train}} \);
3: Select all images \( I_u \) from \textit{COCO}$_{\text{val}}$ which contain \( C_u \) classes objects to construct testing set \( D_{\text{test}} \);
4: return \( D_{\text{train}} \) and \( D_{\text{test}} \);

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.


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
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.