MOST: Multiple Object localization with Self-supervised Transformers for object discovery

Contents

1. Effect of patch size

Due to its dense pixel wise outputs, we analyze the effect of patch size on unsupervised saliency detection in Table 1. We observe that a smaller patch size improves the F1 score but negatively effects the Jaccard index and accuracy for saliency detection.

2. Effect of kernel sizes

In Table 2, we show the effect of the kernel sizes in the pooling layers on the CorLoc performance on PASCAL VOC 2007 dataset. We observe that after kernel size 2, the performance of MOST is robust to different filter sizes. Due to longer runtimes, we use five filters of sizes $k = 1 - 5$ for all the experiments.

3. Class agnostic object detection

Next, we evaluate MOST on the task of category agnostic object detection and report results in Table 3. We use the regions obtained from MOST on the VOC 2007, 2012 trainval sets and COCO20k train set as supervision to train class agnostic object detectors and report results on VOC 2007 trainval, VOC 2007 test, VOC 2012 trainval and COCO minival splits. We use DINO's pretrained weights as initialization to train the object detectors with same hyper-parameters as [1]. We observe consistent improvements across all training and validation sets. When trained on VOC2007, MOST improves upon LOST by 0.91/2 AP/AP_{50} points averaged across all the validation sets. On VOC 2012, MOST improves upon LOST by 0.38/0.69 points across all validation sets. Margin of improvement is larger when trained on COCO20k, a more cluttered and complex dataset than VOC 2007 and 2012, where it improves upon LOST by 1.30/2.68 mAP points across all the validation sets.

4. Experiments on COCO dataset

We train a Faster R-CNN style class agnostic (CAD) and class aware (OD) detectors on the 80k images of COCO 2014 train images and report results in Table 4 and 5 respectively. We use the DINO [2] Resnet [3] 50 weights as initialization and train the detector for 48000 iterations with a batch size of 16 and an initial learning rate of 0.02 on 8 gpus with SynchBatchNorm. The learning rate is dropped at 36000 and 44000 iterations respectively. We add an extra BatchNorm layer for the ROI head after conv5, i.e. Res5ROIHeadsExtraNorm layer in detectron2. To the best of our knowledge, ours is the first work to report results on COCO 2014 train set.

5. Qualitative results

In this section, we show additional qualitative results.

5.1. Object Localization

We show the results of object localization using MOST on PASCAL-VOC 2007, 2012 and COCO20k datasets in Fig. 1-3 respectively. In contrast to recent transformer based object localization and discovery methods [1, 4], MOST can localize multiple objects per image and can do so, without a single round of training. MOST has the potential to localize objects like poles, windows, wall arts and bulletin boards etc. which are typically not in the vocabulary of common object detection datasets [5, 6].

Table 1: Impact of patch size on unsupervised saliency detection on ECSSD, DUTS and DUT-OMRON datasets

Backbone	ECSSD			DUTS			DUT-OMRON		
	max F_β	IoU	Acc (%)	max F_β	IoU	Acc (%)	max F_β	IoU	Acc (%)
ViT-S/8	82.1	59.5	88.2	71.7	63.1	89.4	58.4	44.5	86.9
ViT-S/16	79.1	63.1	89.0	66.6	53.8	89.7	57.0	47.5	87.0

Table 2: Effect on kernel size on CorLoc on VOC07 trainval. Column header [a-b;c] - interval [a,b] with increments of c

Filter size	1	[1-2;1]	[1-3;1]	[1-4;1]	[1-5;1]	[1-7;2]	[2-8;2]	[1-9;2]
CorLoc	72.50	72.50	74.82	74.78	**74.84**	74.50	74.74	74.82

Table 4: Results of CAD trained on COCO 2014 train split. Results reported in mAP

Train →	COCO			
Test →	VOC07 trainval	VOC07 test	VOC12 trainval	COCO minival
AP	12.55	12.81	14.23	4.65
AP_{50}	31.42	32.11	35.49	11.42

Table 5: Results of OD trained on COCO 2014 train split. Results reported in mAP

Train →	COCO		
Clusters →	80	90	100
AP	3.89	3.72	3.90
AP_{50}	9.23	8.98	9.47

5.2. Effect of clusters

In Figures 4, 5 we show the bounding box generated from a cluster, i.e. *pool* of MOST on PASCAL VOC 2007 and COCO20k datasets respectively. MOST automatically identifies the number of clusters, each of which identifies an object and then localizes them without human intervention.

5.3. Effect of patch size

In this section, we compare the outputs of MOST using ViT-S/16 [7] and ViT-S/8 [7] backbones. In Fig. 6, each row consists of three pairs of images. In each pair, the left and right images show the output of MOST using a ViT-S/16 and ViT-S/8 backbone respectively. From row-1,2 we can see that MOST with ViT-S/8 backbone can localize smaller objects which were missed by the backbone with a larger patch size. This comes at the cost of noisier outputs as shown in row-3 of Fig. 6.

5.4. Saliency Detection

MOST can easily be extended for the task of unsupervised saliency detection. We choose the object identified by the largest *pool* as the salient object and demonstrate results on ECSSD [8], DUTS [9] and DUT-OMRON [10] datasets in Figures 8-10 respectively. Each row shows two examples of input and the output of MOST. In each example, the first image is the input, the second image is the mask generated using the largest *pool*, i.e. the output. The third image is the output mask when all the *pools* are used and the fourth image is the ground truth. When only one salient object exists in the input (row-1 of Fig. 8-10) using all the *pools* results in segmenting non salient objects. In the presence of multiple instances of the salient object (row-2 of Fig. 8-10), picking the largest *pool* results in segmenting only a single instance. Finally, in row-3 of Fig. 8-10, we show some failure cases of MOST. Since all the three datasets consists of a majority of images with a single instance, we choose the the mask generated from the largest *pool* as our output.

Table 3: **Results on class-agnostic object detection**: Comparison of MOST with recent works on class-agnostic object detection. We train Faster R-CNN models on VOC 2007, 2012 trainval and COCO20k train splits and report results on VOC2007, VOC2012 trainval, VOC2007 test and COCOminival splits.

Metric	Train →	VOC 2007				VOC 2012				COCO20k			
	Test →	VOC07 trainval	VOC07 test	VOC12 trainval	COCO minival	VOC07 trainval	VOC07 test	VOC12 trainval	COCO minival	VOC07 trainval	VOC07 test	VOC12 trainval	COCO minival
AP	LOST [1]	9.38	10.28	10.00	2.33	11.09	10.86	12.52	2.95	9.69	9.47	10.73	3.10
	MOST	**10.72**	**10.78**	**11.39**	**2.73**	**11.51**	**11.20**	**13.20**	**3.03**	**11.31**	**10.94**	**12.22**	**3.71**
AP_{50}	LOST [1]	27.30	27.22	28.67	7.05	30.26	29.18	33.34	8.55	27.52	26.31	30.17	8.96
	MOST	**29.40**	**28.63**	**31.95**	**7.22**	**31.04**	**29.79**	**34.32**	**8.95**	**30.72**	**29.45**	**33.01**	**10.51**

5.5. Object detection

In Fig. 7 (left) we show the result of object discovery using the output of MOST on VOC 2007 test set. We use K-Means to cluster the regions into 20 clusters and train a Faster R-CNN [11] style detector. Similarly, in Fig. 7 (right) we show the results a Faster RCNN detector with 80 clusters on COCO minival set.

References

[1] Oriane Siméoni, Gilles Puy, Huy V. Vo, Simon Roburin, Spyros Gidaris, Andrei Bursuc, Patrick Pérez, Renaud Marlet, and Jean Ponce. Localizing objects with self-supervised transformers and no labels. In *Proceedings of the British Machine Vision Conference (BMVC)*, November 2021. 1, 2

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 1

[3] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 1

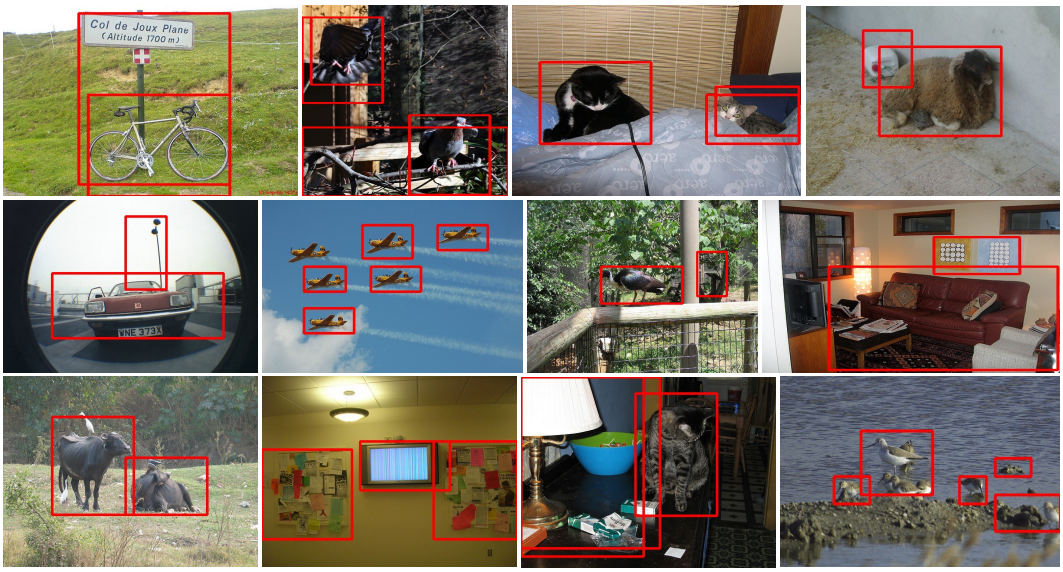

Figure 1: Object localization on PASCAL-VOC 2007.

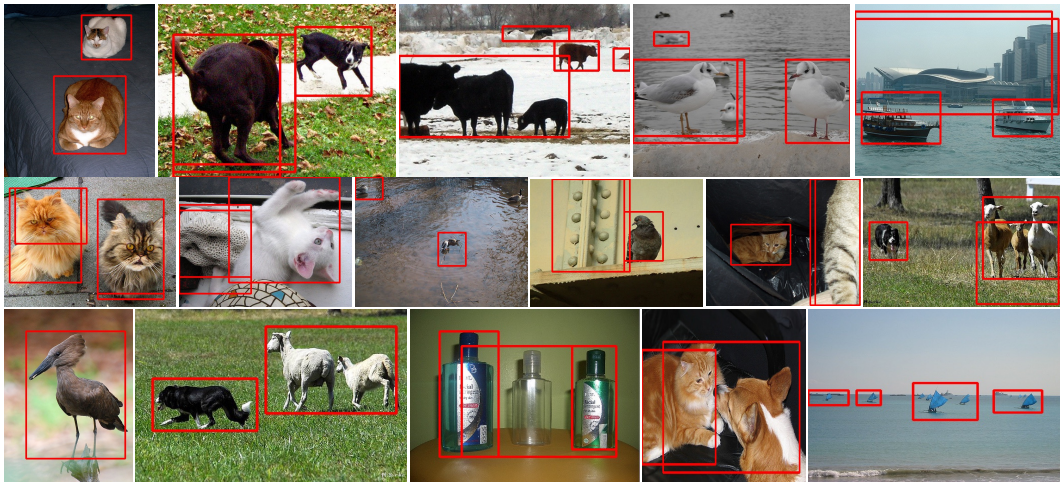

Figure 2: Object localization on PASCAL-VOC 2012.

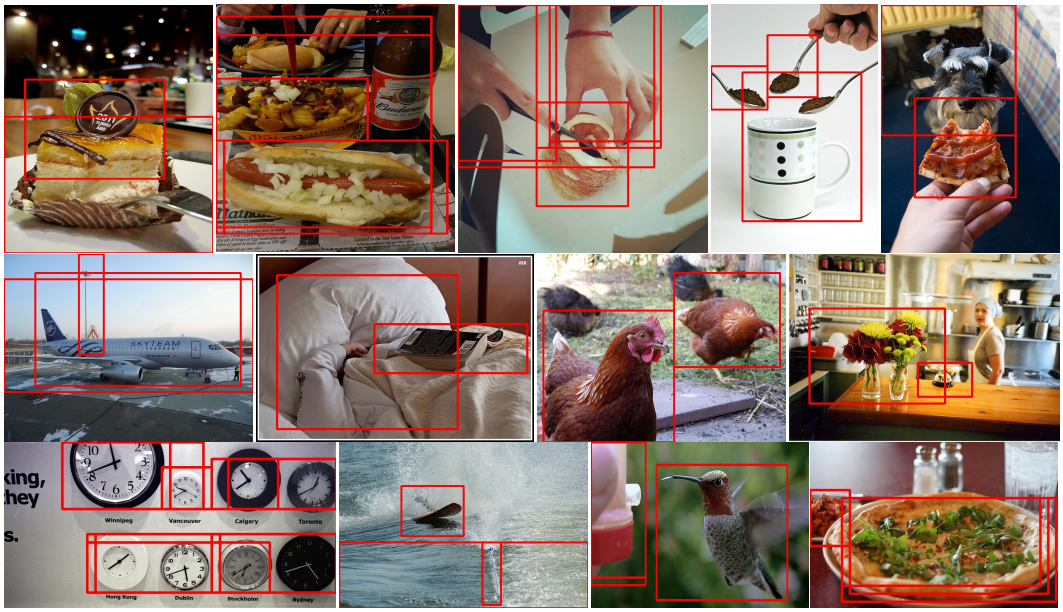

Figure 3: Object localization on COCO20K.

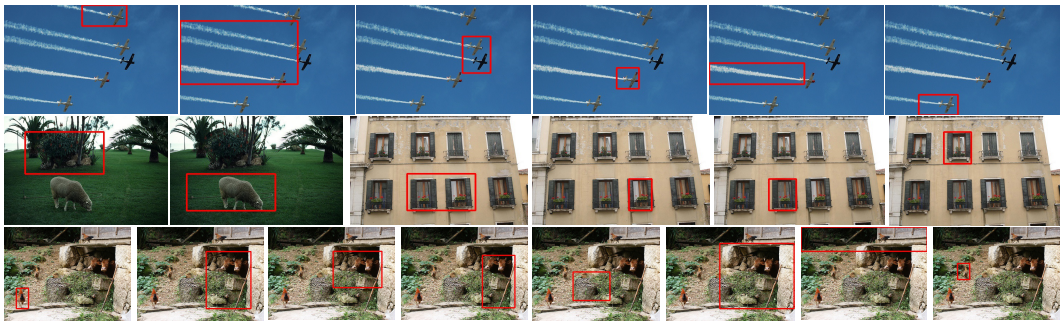

Figure 4: Results on the effect of clusters on PASCAL-VOC 2007.

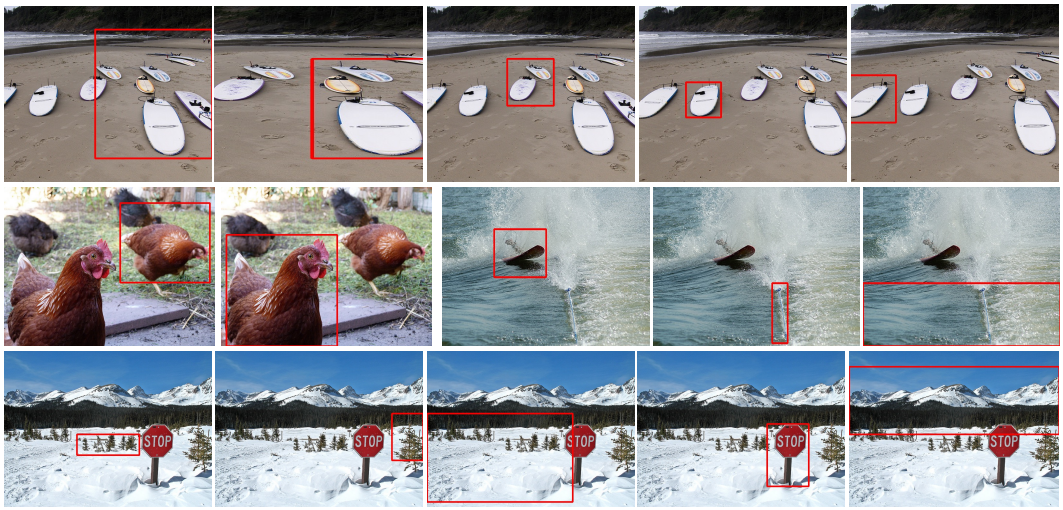

Figure 5: Results on the effect of clusters on COCO20k.

Figure 6: **Effect of patch size:** Each row illustrates three pairs of images showing results of MOST using ViT-S/16 (left) and ViT-S/8 backbones respectively. MOST with ViT-S/8 backbone can localize smaller objects but results in noisier outputs.

[4] Yangtao Wang, Xi Shen, Shell Xu Hu, Yuan Yuan, James Crowley, and Dominique Vaufreydaz. Self-supervised transformers for unsupervised object discovery using normalized cut. In *Conference on Computer Vision and Pattern Recognition*, 2022. 1

[5] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing. 1

[6] Agrim Gupta, Piotr Dollar, and Ross Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 1

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is

Figure 7: **Multi-object discovery(MOST + OD)**. Predictions performed by the class-aware detector on COCO minival (left) and VOC07 test (right). Each class is denoted with a different color.)

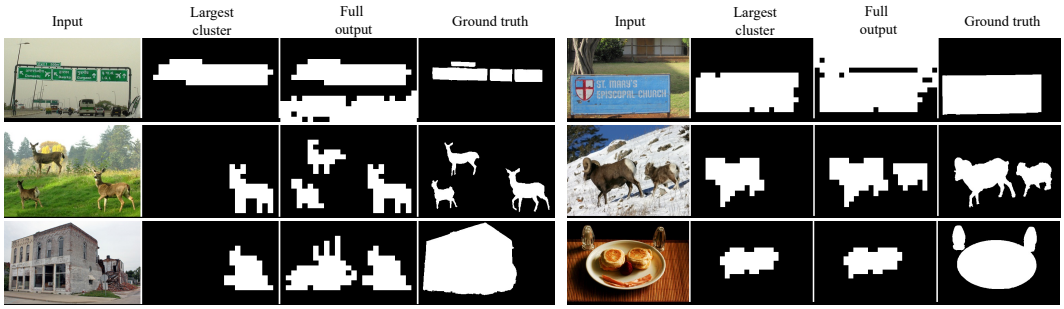

Figure 8: Unsupervised saliency detection on ECSSD [8] dataset using MOST.

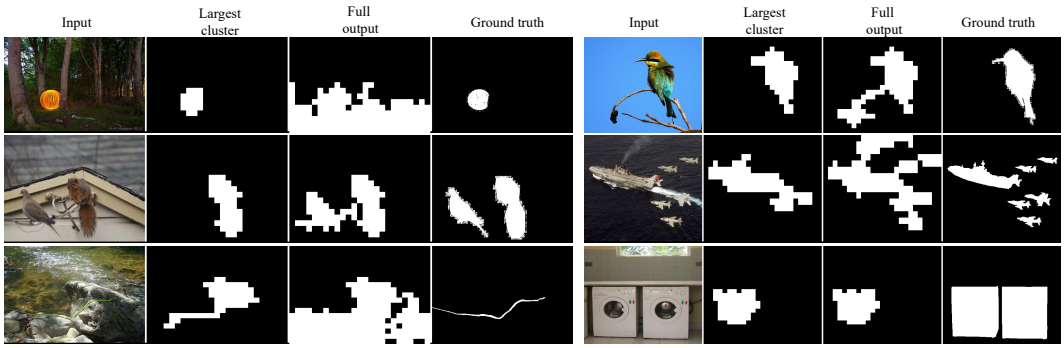

Figure 9: Unsupervised saliency detection on DUTS [9] dataset using MOST.

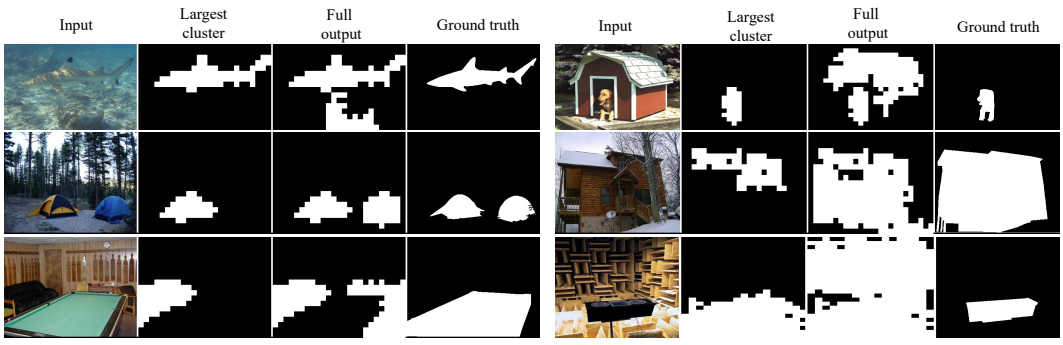

Figure 10: Unsupervised saliency detection on DUT-OMRON [10] using MOST.

worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 2

[8] Jianping Shi, Qiong Yan, Li Xu, and Jiaya Jia. Hierarchical image saliency detection on extended cssd. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38:717–729, 2016. 2, 5

[9] Wang Lijun, Lu Huchuan, Wang Yifan, Feng Mengyang, Wang Dong, Yin Baocai, and Ruan Xiang. Learning to detect salient objects with image-level supervision. In *CVPR*, 2017. 2, 5

[10] Yang Chuan, Zhang Lihe, Lu Huchuan Ruan Xiang, and Yang Ming-Hsuan. Saliency detection via graph-based manifold ranking. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3166–3173. IEEE, 2013. 2, 5

[11] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 91–99. Curran Associates, Inc., 2015. 2