Unsupervised Salient Object Detection with Spectral Cluster Voting

Gyungin Shin[1], Samuel Albanie[2], Weidi Xie[1,3]

[1]Visual Geometry Group, Department of Engineering Science University of Oxford, UK

[2]Department of Engineering, University of Cambridge, UK

[3]Cooperative Medianet Innovation Center, Shanghai Jiao Tong University, China

gyungin@robots.ox.ac.uk, sma71@cam.ac.uk, weidi@sjtu.edu.cn code: https://github.com/NoelShin/selfmask

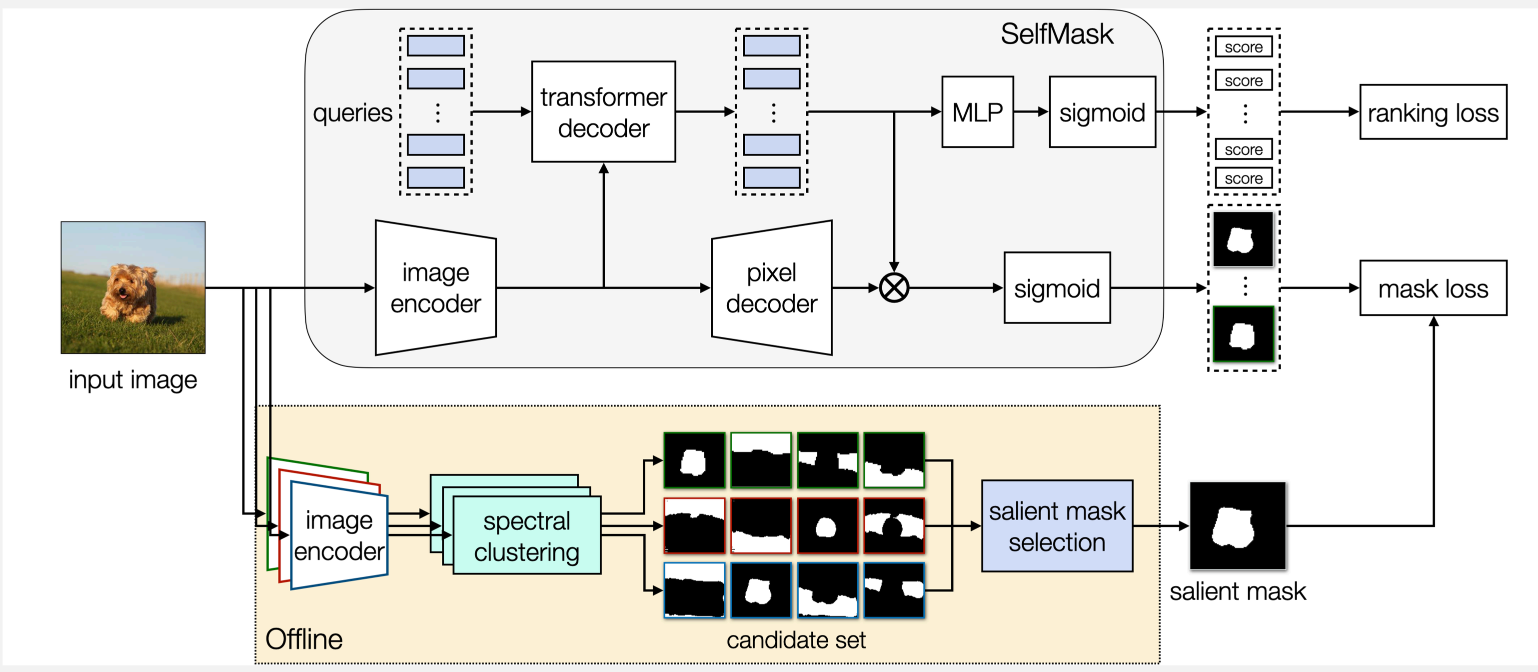
CVPR JUNE 19-24 2022 NEW ORLEANS LOUISIANA

UNIVERSITY OF OXFORD

Overview & contribution

- We revisit spectral clustering and demonstrate its potential for discovering salient objects across various self-supervised features, e.g., MoCov2, SwAV, and DINO.
- Given mask proposals from multiple applications of spectral clustering on different self-supervised features, we pick the most salient mask with a proposed winner-takes-all voting which leverages framing and distinctiveness priors for filtering non-salient masks.
- Using the selected object segmentation as pseudo groundtruth masks, we train a salient object detector, termed SELFMASK, and show that the model outperforms prior approaches on three unsupervised SOD benchmarks.

Proposed method



Given several different self-supervised encoders,

- We first generate a set of pseudo-mask candidates per image using spectral clustering. In the figure, we show 12 masks from clusterings (k=4) on three different encoder features.
- We select the most salient mask among them via the proposed winner-takes-all voting strategy and use it as a pseudo-mask for the image.
- Then we train SELFMASK on the pseudo-masks with two loss functions: a **Mask loss** and a **Ranking loss**. **Mask loss** encourages multiple predictions made by the model to be similar to the pseudo-mask. At the same time, the model is tasked to output an objectness score for each predicted mask such that a score of a prediction closer to the salient mask to be higher via the **Ranking loss**.
- At inference time, we only select a prediction with the highest objectness score.

Experiments

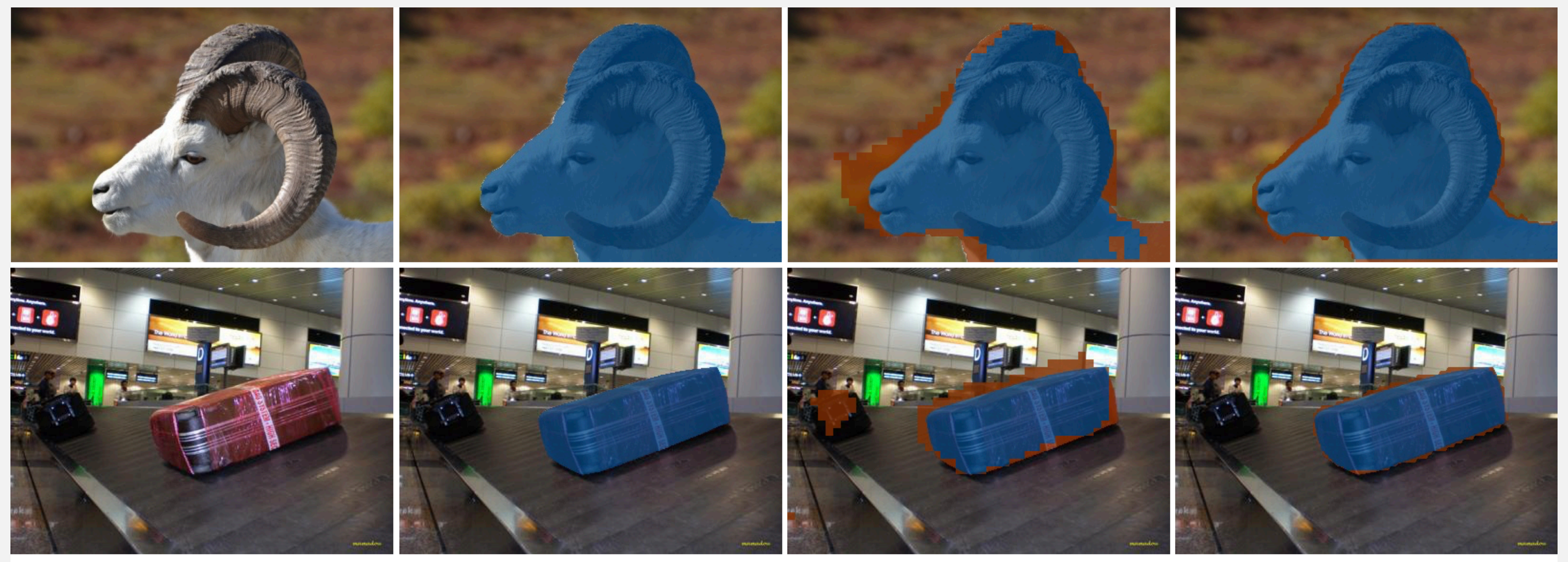


Figure 2: **Sample visualisations of the pseudo-masks and predictions from our model.** Blue and orange coloured regions denote the intersection and difference between a ground-truth and a predicted mask.

Model	Arch.	Cluster.	DUT-OMRON k={2,3,4}	DUTS-TE k={2,3,4}	ECSSD k={2,3,4}
		Convolutional Nets			
MoCov2	ResNet50	k-means	.375	.415	.500
MoCov2	ResNet50	spectral	**.387**	**.454**	**.627**
SwAV	ResNet50	k-means	.399	.444	.542
SwAV	ResNet50	spectral	**.401**	**.458**	**.590**
		Vision Transformer			
DINO	ViT-S/16	k-means	.377	.392	.541
DINO	ViT-S/16	spectral	**.394**	**.417**	**.577**
DINO	ViT-S/8	k-means	.369	.377	.551
DINO	ViT-S/8	spectral	**.398**	**.411**	**.587**

Table 1: **Spectral clustering dominates k-means for self-supervised features.** We report upper bound IoUs to compare the quality of masks produced by k-means and spectral clustering on *self*-supervised features with two different encoder architectures. We report the average of the results from k={2,3,4}.

Features: DINO	MoCov2	SwAV	k	Pseudo-mask	UB
✗	✓	✓	2	.508	.562
			2, 3	.561	.626
			2, 3, 4	.580	.658
✓	✗	✓	2	.473	.553
			2, 3	.538	.644
			2, 3, 4	.559	.682
✓	✓	✗	2	.459	.546
			2, 3	.536	.648
			2, 3, 4	.566	.688
✓	✓	✓	2	.511	.584
			2, 3	.567	.664
			2, 3, 4	**.590**	**.698**

Table 2: **Forming a candidate set with various self-supervised features and multiple k values improves IoU of both pseudo-masks and upper bound masks (UB).** We compare cases with different combinations of self-supervised features and cluster numbers of k=2, $\{2,3\}$ or $\{2,3,4\}$ on HKU-IS.

Selection	Framing prior	HKU-IS	SOD
random	✗	.206	.197
	✓	.464	.277
center	✗	.362	.122
	✓	.442	.392
voting (ours)	✗	.081	.200
	✓	**.590**	**.447**

Table 3: **Winner-takes-all voting and the framing prior both significantly improve mask quality.** We compare our voting strategy to different selection strategies along with the effect of framing prior under the IoU metric. Selection is performed from a candidate set including DINO, MoCov2 and SwAV features with $k = \{2, 3, 4\}$.

Model	DUT-OMRON Acc	IoU	maxF_β	DUTS-TE Acc	IoU	maxF_β	ECSSD Acc	IoU	maxF_β
HS	.843	.433	.561	.826	.369	.504	.847	.508	.673
wCtr	.838	.416	.541	.835	.392	.522	.862	.517	.684
WSC	.865	.387	.523	.862	.384	.528	.852	.498	.683
DeepUSPS	.779	.305	.414	.773	.305	.425	.795	.440	.584
BigBiGAN	.856	.453	.549	.878	.498	.608	.899	.672	.782
E-BigBiGAN	.860	.464	.563	.882	.511	.624	.906	.684	.797
Melas-Kyriazi et al.	.883	.509	-	.893	.528	-	.915	.713	-
LOST	.797	.410	.473	.871	.518	.611	.895	.654	.758
LOST†	.818	.489	.578	.887	.572	.697	.916	.723	.837
TokenCut	.880	.533	.600	.903	.576	.672	.918	.712	.803
TokenCut†	.897	.618	.697	.914	.624	.755	.934	.772	.874
pseudo-masks (Ours)	.811	.403	-	.845	.466	-	.893	.646	-
SELFMASK (Ours)	.901	.582	.680	.923	.626	.750	.944	.781	.889
SELFMASK † (Ours)	**.919**	**.655**	**.852**	**.933**	**.660**	**.882**	**.955**	**.818**	**.956**

Table 4: **Comparison to state-of-the-art unsupervised saliency detection methods on three salient object detection benchmarks.** We observe that SELFMASK yields improved performance over prior state-of-the-art approaches across all benchmarks when trained with the pseudo-masks for the DUTS training images. The best score per column is highlighted in bold. † applies Bilateral solver for post-processing.

Conclusion

- In this work, we address the challenging problem of unsupervised salient object detection (SOD).
- For this, we first observe that self-supervised features exhibit significantly greater object segmentation potential with spectral clustering than with k-means.
- Inspired by this observation, we extract foreground regions among multiple masks generated from differenet self-supervised features, and varying cluster numbers based on the winner-takes-all voting.
- By using the selected masks as pseudo-masks, we train a saliency detection network and show promising results compared to previous unsupervised methods on various SOD benchmarks.