Boosting Unsupervised Segmentation Learning

Alp Eren Sari
University of Bern
Switzerland
alp.sari@unibe.ch

Francesco Locatello
Institute of Science and Technology Austria
Austria
francesco.locatello@ista.ac.at

Paolo Favaro
University of Bern
Switzerland
paolo.favaro@unibe.ch

Abstract

We present two practical improvement techniques for unsupervised segmentation learning. These techniques address limitations in the resolution and accuracy of predicted segmentation maps of recent state-of-the-art methods. Firstly, we leverage image post-processing techniques such as guided filtering to refine the output masks, improving accuracy while avoiding substantial computational costs. Secondly, we introduce a multi-scale consistency criterion, based on a teacher-student training scheme. This criterion matches segmentation masks predicted from regions of the input image extracted at different resolutions to each other. Experimental results on several benchmarks used in unsupervised segmentation learning demonstrate the effectiveness of our proposed techniques.

1 Introduction

The task of segmenting objects in images is a fundamental first step in extracting semantic information. Current approaches [15, 14] have reached remarkable performance and capabilities by combining various annotation modalities (text, points, bounding boxes, and segmentation masks) and by scaling the training to large datasets. However, recently, several methods [10, 31, 20, 19, 3, 28] have shown that it is possible to learn to segment objects in an unsupervised manner, i.e., without the guidance of any human annotation. Besides the intrinsic scientific significance of such unsupervised learning methods, their main advantage is that they could potentially be scaled to extremely large datasets and across multiple imaging modalities with limited human effort.

Despite encouraging results, these methods seem to be either limited in resolution [19] or to rely on difficult training schemes [3]. Since many of the existing methods [18, 28, 19] rely on DINO features [5], which reduce the initial image resolution by a factor of 8 or 16, we introduce two rather general "tricks" to enhance the resolution of predicted segmentation maps and demonstrate their capabilities on the recent state of the art method *Sempart* for unsupervised segmentation learning [19]. Our first trick is to use post-processing to enhance the resolution of the output masks. We show that well-established image processing methods such as guided filtering [13] can easily improve the accuracy of the predicted segmentation masks when the input image luminance is used as the filter guidance.

Our second trick is a general criterion to enhance the resolution of dense predictions. We introduce a *multi-scale consistency* criterion on the predicted segmentation mask. We do so by using a teacher-student training scheme, where the teacher network takes a zoomed-in region of an image as input

Self-Supervised Learning - Theory and Practice Workshop, 38th Conference on Neural Information Processing Systems (NeurIPS 2024).

and the student network takes the whole image as input. Then, we zoom in on the same region of the predicted segmentation mask from the student network's prediction and match it to the prediction from the teacher. Because we use the output of the teacher network as a target, we only backpropagate through the student network.

In summary, our contributions are

- We achieve new state-of-the-art (SotA) results in unsupervised saliency segmentation on the DUT-OMRON [30], DUTS-TE [27], and ECSSD [22] datasets;
- We introduce two novel and general techniques to enhance the resolution of the segmentation masks predicted by existing SotA that are computationally efficient;
- Our techniques are easy to apply and have been thoroughly tested on several benchmarks and in combination with different methods; the code will be made available upon publication.

2 Related Work

Spectral Methods. Graph representations are commonly utilized in computer vision and computer graphics for partitioning problems. [21] propose the *normalized cuts* method for image segmentation to handle the bias towards small-isolated segments in the minimum graph cut method [29].
Self-supervised representations (SSL). The idea of training deep networks without any supervision to obtain effective image representations gained popularity with the success of [6, 12, 4, 5, 11]. In particular, DINO [5] features based on ViT [7] demonstrate capabilities beyond image classification.
Unsupervised Semantic Segmentation. Several recent approaches, such as [24, 18, 28], explore the use of DINO features and graph cuts for semantically meaningful image segmentation. However, these methods often exhibit high computational complexity and relatively low segmentation accuracy. In contrast, [2, 3] rely on the local "shiftability" of segmented objects, but face challenges due to the complexity of their adversarial training schemes. [17] explores pre-trained generative adversarial networks (GAN). However, it is also challenging to generalize this approach due to the limitations of existing GANs. Other methods, like [26, 25, 23, 19], leverage various strategies, such as push-pull forces between pixels, and pseudo-masks, to distill segmentation networks. However, these approaches often suffer from low-resolution image features, limiting the quality of the generated masks.

3 Two Tricks

We introduce two tricks that can be incorporated into different unsupervised segmentation methods in a flexible manner: guided filtering[13] and random cropping-based equivariance loss.
Guided Filtering. First, the RGB image -whose saliency mask is predicted- is converted to grayscale. Afterward, the guided filter [13] is applied to the predicted segmentation mask using the grayscale image as guidance.
Multi-Scale Consistency. One challenge associated with incorporating a segmentation head over a ViT [7] feature map (as with DINO) is the loss of details, particularly for small objects or structures within the image. Sempart [19] addresses this issue by feeding the input image into the segmentation head and by applying a total variation loss on the resulting high-resolution mask. We aim to tackle this problem by introducing an equivariance loss based on random cropping depicted in Fig. 1. The proposed equivariance loss is computed through the following steps:

1. Given an input image and a segmentation framework denoted as I and $f(\cdot)$ respectively, predict a mask denoted as $\hat{s}$. This process can be expressed as $f(I) = \hat{s}$.
2. Randomly crop the same region of the input image and the predicted mask and enlarge both of them to the original input image size. This operation on I, $\hat{s}$ can be expressed as $\text{CU}(I, \hat{s}) = \{I_c, s_c\}$ where $\text{CU}(\cdot, \cdot)$, I_c, and s_c are the crop-zoom function, the resultant image and mask.
3. Use the resultant image to predict a mask and refine this mask with guided image filtering, which can be expressed as $f(I_c) = \hat{s}_c$.
4. Apply stop gradient afterward to prevent a collapse of the predicted masks, which we denote as $\text{SG}[\text{GF}(\hat{s}_c)] = s_{target}$, where s_{target} is the final target for our loss, and $\text{SG}[\cdot]$ is the stop gradient operation.

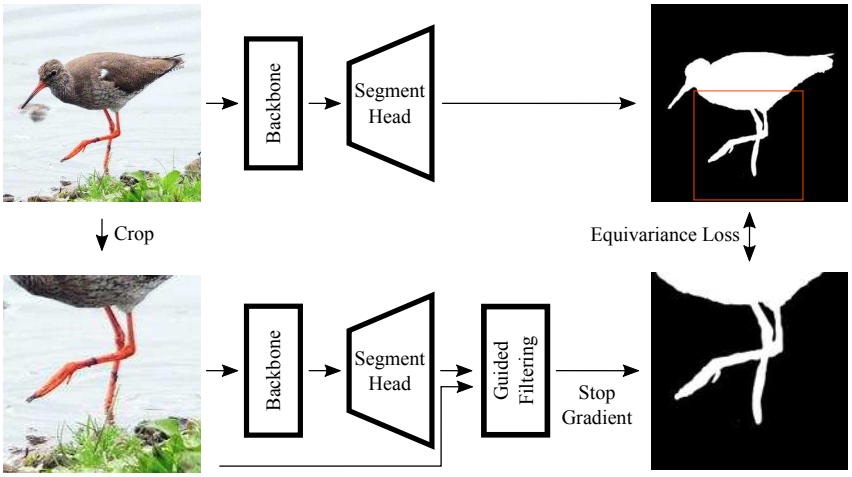

Figure 1: The illustration of the proposed multi-scale consistency procedure. First, we predict a segmentation from an input image. Then, we randomly crop a portion of the input image, predict a more detailed segmentation mask, refine this prediction with guided filtering, and apply a stop-gradient operation to the final mask to prevent a mask prediction collapse. Finally, we calculate the mean squared error between the corresponding region of the initial mask and the detailed target mask.

Table 1: The unsupervised saliency segmentation results on DUT-OMRON [30], DUTS-TE [27], and ECSSD [22]. The best results are indicated in **bold**. Sempart* indicates our implementation of Sempart [19] since the code is currently unavailable. "Ours" indicates the application of the proposed trick over the Sempart baseline. We demonstrate improvements over the baseline implementation we have.

	DUT-OMRON [30]			DUTS-TE [27]			ECSSD [22]		
Model	Acc	IoU	$maxF_\beta$	Acc	IoU	$maxF_\beta$	Acc	IoU	$maxF_\beta$
LOST [24]	.797	.410	.473	.871	.518	.611	.895	.654	.758
TokenCut [28]	.880	.533	.600	.903	.576	.672	.918	.712	.803
FOUND [25] - single	.920	.586	.683	.939	.637	.733	.912	.793	.946
MOVE [3]	.923	.615	.712	.950	.713	.815	.954	.830	.916
Sempart [19]	**.932**	.668	.764	**.959**	.749	.867	**.964**	.855	**.947**
Sempart*	.908	.664	.758	.949	.761	.856	.950	.853	.925
Ours	.920	**.684**	**.765**	.956	**.787**	**.869**	.952	**.862**	.931
MOVE [3] + BF	.931	.636	.734	.951	.687	.821	.953	.801	.916
Sempart + BF	**.942**	**.698**	**.799**	**.958**	**.749**	**.879**	**.963**	**.850**	**.944**
Sempart* + BF	.914	.689	.770	.950	.765	.857	.950	.849	.925
Ours + BF	.920	.675	.763	.950	.742	.855	.945	.830	.920
SELFMASK w/ MOVE	.933	.666	.756	.954	.728	.829	.956	.835	.921
SELFMASK w/ Sempart	**.942**	**.698**	**.799**	**.958**	**.749**	.879	**.963**	.850	.944
SELFMASK w/ Sempart*	.923	.674	.768	.953	.743	.866	.962	**.854**	.942
SELFMASK w/ Ours	.935	.685	.790	.956	.747	**.881**	.962	.852	**.948**

5. Calculate the mean squared error between the new prediction and the corresponding region of the first segmentation prediction as $\mathcal{L}_{eq} = \frac{1}{hw}\|s_c - s_{target}\|_2^2$ where h and w are the height and width of the corresponding masks $\hat{s}_c$ and s_{target}.

In our experiments, we added $\mathcal{L}_{eq}$ multiplied with the parameter λ_{eq} to the overall loss of Sempart [19]. We also used the cropped image to train the network.

4 Experimental Results

We evaluate the tricks we propose on Sempart [19] since it is the current state-of-the-art unsupervised segmentation learning method, it trains stably, and it converges quickly. The evaluations are performed

Table 2: Comparison of single object detection on Pascal VOC07 [8], Pascal VOC12 [9], COCO20K [16] with CorLoc metric. The best results are indicated in **bold**. Sempart* indicates our implementation of Sempart [19]. ($\uparrow z$) indicate the improvements over the baseline.

Model	VOC07 [8]	VOC12 [9]	COCO20K [16]
LOST [24]	61.9	64.0	50.7
Deep Spectral [18]	62.7	66.4	50.7
TokenCut [28]	68.8	72.1	58.8
MOVE [3]	**76.0**	**78.8**	**66.6**
Sempart [19]	75.1	76.8	66.4
Sempart*	71.7	75.9	61.3
Ours	74.4 ($\uparrow$ 2.7)	77.0 ($\uparrow$ 1.1)	63.4 ($\uparrow$ 2.1)

for saliency segmentation and single object detection since Sempart [19] and previous works were evaluated on these tasks. All the evaluations were performed using the fine branch of Sempart [19]. The improved baseline with the proposed tricks is denoted as "ours".

4.1 Unsupervised Saliency Segmentation

Datasets. We used the DUTS-TR [27] for training in every experiment, as [23, 3, 19]. To evaluate the saliency segmentation, we used DUT-OMRON [30], DUTS-TE [27], and ECSSD [22] datasets.
Evaluation. Three metrics are reported in our work as [23, 3, 19], which are pixel mask accuracy (Acc), intersection over union (IoU), and $\max F_\beta$ [28].
Results. In Table 1, we present a comprehensive comparison of our results with the recent state-of-the-art method Sempart [19]. Our implementation, denoted as "Sempart*", undergoes evaluation across three distinct criteria: baseline results, post-processed baseline results employing bilateral filtering [1], and outcomes derived from SELFMASK [23], trained with pseudo-masks generated by various methods. The comparison reveals the efficacy of our approach in advancing unsupervised saliency segmentation.

Our results show state-of-the-art Intersection over Union (IoU) performance across all three datasets. When considering the notable gaps between our implemented baseline and the reported baseline, particularly in terms of accuracy (Acc) and $\max F_\beta$ metrics, it becomes evident that our approach has the potential to establish a new state-of-the-art benchmark with precisely the same baseline.

4.2 Single Object Detection

Datasets. The evaluation of single object detection was done on three datasets, which are the train split of COCO20K [16], the training and validation splits of Pascal VOC07 [8], and Pascal VOC12 [9].
Evaluation. We follow the evaluation protocol of the previous methods [28, 3, 19] and report *Correct Localization* (CorLoc) metric.
Results. In Table 2, we conduct a comprehensive comparison between our results and the recent state-of-the-art MOVE [3], as well as our baseline in object detection. Our tricks significantly enhance the baseline across all three datasets considered. This improvement is evident in detection accuracy, underscoring the effectiveness of our approach in advancing unsupervised object detection.

5 Conclusion

We introduce two practical methods that can be flexibly incorporated into unsupervised image segmentation. We demonstrate their effectiveness with state-of-the-art results on unsupervised saliency segmentation tasks, combined with the recent state-of-the-art baseline. Moreover, in the appendix, we demonstrate that these two methods can be effective even if the baseline was trained with different backbones and combined with other recent unsupervised segmentation methods. We also include related background information, more qualitative results illustrating the gains of our methods, and an ablation study in unsupervised saliency segmentation and unsupervised object discovery in the appendix.

References

[1] Jonathan T Barron and Ben Poole. The fast bilateral solver. In *ECCV*, pages 617–632. Springer, 2016.

[2] Adam Bielski and Paolo Favaro. Emergence of object segmentation in perturbed generative models. *NeurIPS*, 32, 2019.

[3] Adam Bielski and Paolo Favaro. Move: Unsupervised movable object segmentation and detection. *NeurIPS*, 35:33371–33386, 2022.

[4] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *NeurIPS*, 33:9912–9924, 2020.

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, pages 9650–9660, 2021.

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607. PMLR, 2020.

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2020.

[8] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88:303–338, 2010.

[9] Mark Everingham and John Winn. The pascal visual object classes challenge 2012 (voc2012) development kit. *Pattern Anal. Stat. Model. Comput. Learn., Tech. Rep*, 2007(1-45):5, 2012.

[10] Mark Hamilton, Zhoutong Zhang, Bharath Hariharan, Noah Snavely, and William T Freeman. Unsupervised semantic segmentation by distilling feature correspondences. In *ICLR*, 2021.

[11] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, pages 16000–16009, 2022.

[12] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9729–9738, 2020.

[13] Kaiming He, Jian Sun, and Xiaoou Tang. Guided image filtering. *IEEE TPAMI*, 35(6):1397–1409, 2012.

[14] Lei Ke, Mingqiao Ye, Martin Danelljan, Yu-Wing Tai, Chi-Keung Tang, Fisher Yu, et al. Segment anything in high quality. *NeurIPS*, 36, 2024.

[15] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

[16] 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 *ECCV*, pages 740–755. Springer, 2014.

[17] Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. Finding an unsupervised image segmenter in each of your deep generative models. In *ICLR*, 2021.

[18] Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. Deep spectral methods: A surprisingly strong baseline for unsupervised semantic segmentation and localization. In *CVPR*, pages 8364–8375, 2022.

[19] Sriram Ravindran and Debraj Basu. Sempart: Self-supervised multi-resolution partitioning of image semantics. In *ICCV*, pages 723–733, 2023.

[20] M Seitzer, M Horn, A Zadaianchuk, D Zietlow, T Xiao, C Simon-Gabriel, T He, Z Zhang, B Schölkopf, Thomas Brox, et al. Bridging the gap to real-world object-centric learning. In *ICLR*, 2023.

[21] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE TPAMI*, 22(8):888–905, 2000.

[22] Jianping Shi, Qiong Yan, Li Xu, and Jiaya Jia. Hierarchical image saliency detection on extended cssd. *IEEE TPAMI*, 38(4):717–729, 2015.

[23] Gyungin Shin, Samuel Albanie, and Weidi Xie. Unsupervised salient object detection with spectral cluster voting. In *CVPR*, pages 3971–3980, 2022.

[24] 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 *BMVC*, 2021.

[25] Oriane Siméoni, Chloé Sekkat, Gilles Puy, Antonín Vobeckỳ, Éloi Zablocki, and Patrick Pérez. Unsupervised object localization: Observing the background to discover objects. In *CVPR*, pages 3176–3186, 2023.

[26] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, and Luc Van Gool. Unsupervised semantic segmentation by contrasting object mask proposals. In *ICCV*, pages 10052–10062, 2021.

[27] Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to detect salient objects with image-level supervision. In *CVPR*, pages 136–145, 2017.

[28] Yangtao Wang, Xi Shen, Yuan Yuan, Yuming Du, Maomao Li, Shell Xu Hu, James L Crowley, and Dominique Vaufreydaz. Tokencut: Segmenting objects in images and videos with self-supervised transformer and normalized cut. *IEEE TPAMI*, 2023.

[29] Zhenyu Wu and Richard Leahy. An optimal graph theoretic approach to data clustering: Theory and its application to image segmentation. *IEEE TPAMI*, 15(11):1101–1113, 1993.

[30] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via graph-based manifold ranking. In *CVPR*, pages 3166–3173, 2013.

[31] Andrii Zadaianchuk, Matthaeus Kleindessner, Yi Zhu, Francesco Locatello, and Thomas Brox. Unsupervised semantic segmentation with self-supervised object-centric representations. In *ICLR*, 2022.

[32] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. Image bert pre-training with online tokenizer. In *ICLR*, 2021.

A Additional Background

In this section we briefly revise some basic definitions to make the paper self-contained and also to provide a context for our proposed techniques.

Normalized Cuts. Normalized cuts [21] is a method to partition a weighted undirected graph $G = (V, E)$ with weights w_{ij} representing the similarity between nodes v_i and v_j, into partitions A and B by minimizing the following loss

$$\text{Ncut}(A, B) = \frac{\text{cut}(A, B)}{\text{assoc}(A, V)} + \frac{\text{cut}(B, A)}{\text{assoc}(B, V)}, \tag{1}$$

$$\text{where cut}(A, B) \doteq \sum_{u \in A, v \in B} w_{uv} \quad \text{and} \quad \text{assoc}(A, V) \doteq \sum_{u \in A, t \in V} w_{ut}.$$

$\text{cut}(B, A)$, and $\text{assoc}(B, V)$ are also defined similarly. By minimizing Ncut with respect to A and B, the connectivity between partitions A and B is minimized and the connectivity within the partitions

is maximized. The weights are collected in an *affinity matrix* W such that its i, j-th entry $W_{i,j} = w_{ij}$. The minimization of normalized cuts loss is NP-complete; hence, [21] proposed a relaxation of the proposed loss which leads to a generalized eigenvalue problem.

TokenCut. [5] shows that the attention maps in DINO strongly relate to the location and shape of the object present in the input image. Thus, TokenCut [28] proposes to utilize DINO features for semantic segmentation combined with normalized cuts as a *post-processing* step. TokenCut [28] minimizes the normalized cuts' loss with the following affinity matrix

$$w_{ij} = \begin{cases} 1 \mid \langle F_{v_i}, F_{v_j} \rangle > \tau \\ \epsilon \mid \text{otherwise.} \end{cases} \tag{2}$$

where F_{v_i} is the DINO feature corresponding to the node v_i, and τ and ϵ are scalar thresholds to be tuned.

Sempart. Sempart [19] instead leverages DINO features by training a convolutional neural networks (CNN) segmentation head to minimize a loss based on the normalized cut cost. The segmentation head generates two segmentation predictions: one, $S_{\text{coarse}} \in [0,1]^{|V|}$, at a low-resolution and one, $S_{\text{fine}} \in [0,1]^{HW}$, at high-resolution, where $|V|$ is the number of DINO features, and H and W are the height and width of the input image. S_{coarse} minimizes the following variation of the normalized cuts' loss

$$\mathcal{L}_{\text{Ncut}} = \text{Ncut}(A, B) = \frac{S^T W(1-S)}{S^T W \mathbf{1}} + \frac{(1-S)^T W S}{(1-S)^T W \mathbf{1}}, \tag{3}$$

where $\mathbf{1}$ is a vector of ones. Eq. 3 is optimized only for the low-resolution mask S_{coarse}. High-resolution mask S_{fine} is guided through a consistency loss

$$\mathcal{L}_{SR} = \|\text{down}(S_{\text{fine}}) - S_{\text{coarse}}\|_2^2, \tag{4}$$

which matches the downsampled S_{fine} and S_{coarse}, and $\mathcal{L}_{\text{GTV-fine}}$ is a graph total variation (GTV) loss that works on the high-resolution mask. An additional GTV loss $\mathcal{L}_{\text{GTV-coarse}}$ computed with S_{coarse} is also used for additional guidance. $\mathcal{L}_{\text{GTV-fine}}$ and $\mathcal{L}_{\text{GTV-coarse}}$ are given as

$$\mathcal{L}_{\text{GTV-fine/coarse}} = \frac{1}{2} \sum_{(i,j) \in E} a_{ij}(s_i - s_j)^2, \tag{5}$$

$$a_{ij}^{\text{fine}} = \exp(-\|x_i - x_j\|_2^2 / \sigma), \tag{6}$$

$$a_{ij}^{\text{coarse}} = w_{ij} \mathbf{I}\{i \in \mathcal{N}(j)\}, \tag{7}$$

where $\mathcal{N}(j)$ is set of nodes adjacent to node j, and $\mathbf{I}$ is the indicator function. The only difference between $\mathcal{L}_{\text{GTV-fine}}$ and $\mathcal{L}_{\text{GTV-coarse}}$ is that x_i is a pixel value from the input image for fine GTV, but it is the DINO feature for the coarse GTV. Finally, the segmentation head is trained with the combined loss

$$\mathcal{L}_{\text{sempart}} = \mathcal{L}_{\text{Ncut}} + \lambda_{\text{GTV-coarse}} \mathcal{L}_{\text{GTV-coarse}} + \lambda_{\text{GTV-fine}} \mathcal{L}_{\text{GTV-fine}} + \lambda_{\text{SR}} \mathcal{L}_{\text{SR}}. \tag{8}$$

B Additional Results

B.1 Unsupervised Saliency Segmentation

We provide some qualitative results in Fig. 2 for the cases where the segmentation is improved, unaffected, and degraded.

B.2 Single Object Detection

We share the implementation details of the single object detection evaluation and some qualitative results here to illustrate how our tricks boost object detection. Specifically, Fig. 3 visually depicts how random cropping facilitates the detection of small objects and, in some instances, objects with occlusion. **Evaluation Details.** First, the predicted mask is split into parts if it's not fully connected. Then, we select the mask with the largest bounding box as our object prediction and calculate the IoU between our prediction and the ground-truth bounding box. Finally, the percentage of the IoUs higher than 0.5 is calculated, known as the *Correct Localization* (CorLoc) metric.

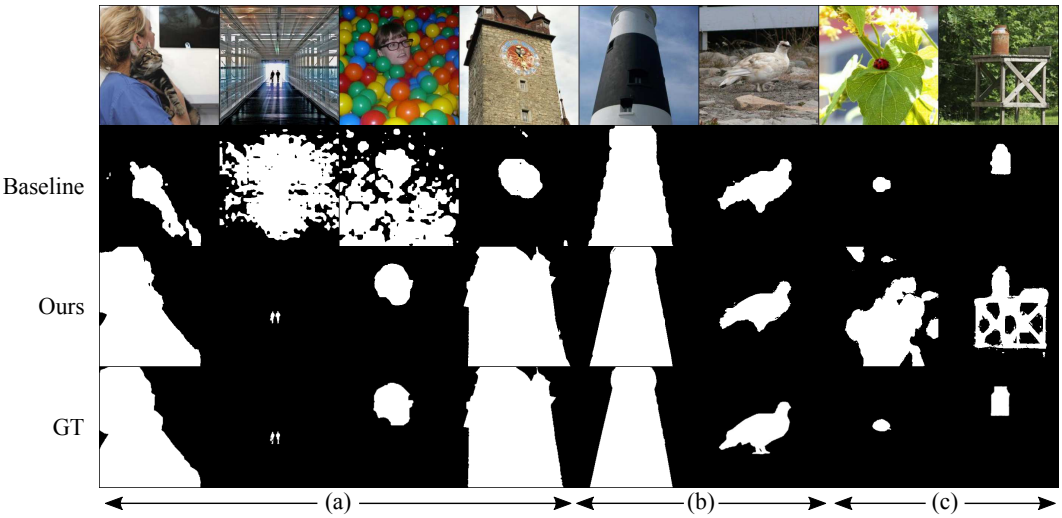

Figure 2: A qualitative comparison between the baseline and the improved segmentation with our proposed tricks (Ours). The first row shows images sampled from DUTS-TE [27], the second row shows the baseline segmentation predictions, the third row shows the segmentation results with our tricks, and the last row shows the ground truth segmentation masks. In (a) we show 4 image samples where we achieve a significant improvement over the baseline; in (b) we show two image samples where we achieve the same results as in the baseline; in (c) we show two image samples where the tricks make the segmentation masks worse (relative to the ground truth mask). We point out that some of the incorrect masks may also be due to the inherent ambiguity of the unsupervised segmentation task.

B.3 Ablations

We conducted various ablation experiments to demonstrate the effectiveness of the proposed enhancements. These experiments involved evaluating the impact of the proposed techniques compared to the baseline in unsupervised saliency segmentation and single object detection. Additionally, we performed an ablation study on the guided filtering, comparing its performance with recent methods such as TokenCut [28] and MOVE [3] in unsupervised saliency segmentation. For the saliency segmentation ablation, we trained different baselines using various SSL backbones and demonstrated that our techniques consistently outperformed the baselines across different backbones.
The tricks on saliency segmentation. In Table 3, we present the Intersection over Union (IoU) results obtained from the evaluation of unsupervised saliency segmentation on three datasets: DUT-OMRON [30], DUTS-TE [27], and ECSSD [22]. The original baseline is trained with DINO [5] ViT-S/8 as the backbone. Additionally, we provide results for baselines trained with DINO ViT-S/16 and IBOT [32] ViT-S/16 as the backbones. Our findings indicate that across different backbones, the proposed techniques significantly enhance segmentation performance. Notably, guided filtering emerges as a key contributor to the improved segmentation results.

The tricks on single object detection. In Table 4, we demonstrate the enhancement achieved by our techniques in object detection based on the CorLoc metric. The evaluation is performed on datasets comprising the train and validation splits of Pascal VOC07 [8] and Pascal VOC12 [9], as well as the train split of COCO20K [16]. Notably, unlike the case of saliency segmentation, we observe that random cropping yields the most significant improvement in object detection performance.
Guided filtering with different methods. In Table 5, our findings indicate that the application of guided filtering significantly enhances the performance of recent saliency segmentation methods TokenCut [28] and MOVE [3]. We present results Acc, IoU, and the $maxF_\beta$ metric across evaluation datasets including DUT-OMRON [30], DUTS-TE [27], and ECSSD [22]. Comparative analysis with baseline methods, incorporating bilateral filtering [1], is also included. Notably, the results in Table 5 demonstrate consistent improvement in all metrics with guided filtering, which contrasts with the performance of bilateral filtering.

Baseline Ours GT

Figure 3: Some examples cases from Pascal VOC12 [9] where our tricks improve object detection over baseline.

Table 3: The ablations on unsupervised saliency segmentation with mIoU on three datasets; DUT-OMRON [30], DUTS-TE [27], and ECSSD [22]. The best results are indicated with **bold**.

Backbone	Crop	GF	DUT-OMRON [30]	DUTS-TE [27]	ECSSD [22]
DINO [5] ViT-S/8	×	×	.664	.761	.853
	✓	×	.670	.772	.852
	×	✓	**.684**	.784	**.869**
	✓	✓	**.684**	**.787**	.862
DINO [5] ViT-S/16	×	×	.604	.692	.784
	✓	×	.647	.710	.812
	×	✓	.608	.697	.788
	✓	✓	**.652**	**.714**	**.816**
IBOT [32] ViT-S/16	×	×	.584	.650	.793
	✓	×	.588	.654	.797
	×	✓	.600	.664	.806
	✓	✓	**.604**	**.668**	**.811**

B.4 Limitations

Our method exhibits challenges in scenarios where saliency within an image is unambiguous or when multiple objects are present in the scene. Some visual examples can be seen in Fig. 4, where our method tends to segment all objects in the scene, whereas human annotators are inclined to select smaller and easily movable objects. Moreover, in instances where the salient object and the background share visual similarities our method encounters difficulty, making it challenging to distinguish between them accurately.

Table 4: The ablations on unsupervised single object detection with CorLoc on three datasets; Pascal VOC07 [8], Pascal VOC12 [9], COCO20K[16]. The best results are indicated with **bold**.

Crop	GF	VOC07 [8]	VOC12 [9]	COCO20K [16]
×	×	71.7	75.9	61.3
✓	×	74.2	**77.0**	**64.3**
×	✓	71.7	75.7	61.3
✓	✓	**74.4**	**77.0**	63.4

Table 5: The ablations of guided filtering [13] and bilateral filtering [1] with TokenCut [28] and MOVE [3] on unsupervised saliency segmentation with same datasets and metrics in Table 1. The best results for each method are indicated with **bold**.

Model	DUT-OMRON [30]			DUTS-TE [27]			ECSSD [22]		
	Acc	IoU	$maxF_\beta$	Acc	IoU	$maxF_\beta$	Acc	IoU	$maxF_\beta$
TokenCut	.880	.533	.600	.903	.576	.672	.918	.712	.803
TokenCut + GF	.880	.553	.676	.907	.597	.740	.925	.740	.861
TokenCut + BF	**.897**	**.618**	**.697**	**.914**	**.624**	**.755**	**.934**	**.772**	**.874**
MOVE	.923	.615	.712	.950	.713	.815	.954	.830	.916
MOVE + GF	.925	.615	.631	**.952**	**.726**	.813	**.958**	**.843**	**.923**
MOVE + BF	**.931**	**.636**	**.734**	.951	.687	**.821**	.953	.801	.916

Figure 4: Some of the failure cases of our method from DUTS-TE [27]. In most cases the saliency is unambiguous or the background and the foreground are almost indistinguishable.