SEMPART: Self-supervised Multi-resolution Partitioning of Image Semantics

Sriram Ravindran
Adobe
sravindr@adobe.com

Debraj Basu
Adobe
dbasu@adobe.com

Figure 1 Each image is the original image with an overlayed saliency mask. The first row of every column utilizes the ground truth saliency mask, the second and third rows overlay with the self-supervised SEMPART-coarse and -fine masks on the same image respectively.

Abstract

Accurately determining salient regions of an image is challenging when labeled data is scarce. DINO-based self-supervised approaches have recently leveraged meaningful image semantics captured by patch-wise features for locating foreground objects. Recent methods have also incorporated intuitive priors and demonstrated value in unsupervised methods for object partitioning. In this paper, we propose SEMPART*, which jointly infers coarse and fine bi-partitions over an image's DINO-based semantic graph. Furthermore,* SEMPART *preserves fine boundary details using graph-driven regularization and successfully distills the coarse mask semantics into the fine mask. Our salient object detection and single object localization findings suggest that* SEMPART *produces high-quality masks rapidly without additional post-processing and benefits from co-optimizing the coarse and fine branches.*

1. Introduction

Identifying salient regions of an image prone to holding visual attention remains a long-standing fuzzy problem [59] relying significantly on carefully annotated data [51, 5, 54]. Recently self-supervised (SSL) mechanisms based on large-scale pre-trained backbones [9, 6, 22], such as DINO [7], have demonstrated increased capability in segmenting images [21, 30] and extracting objects in the foreground [41, 39, 54, 4, 42].

The unavailability of labels is limiting to inferring high-quality object masks. However, many recent methods have demonstrated that incorporating well-informed priors into the partitioning process is significantly beneficial to finding saliency regions and foreground objects in an unsupervised setting [36, 41, 46, 47, 31, 54, 4, 39].

Different forms of statistical independence of the foreground have driven recent approaches, with the most recent state-of-the-art focusing on movability [4] of the salient ob-

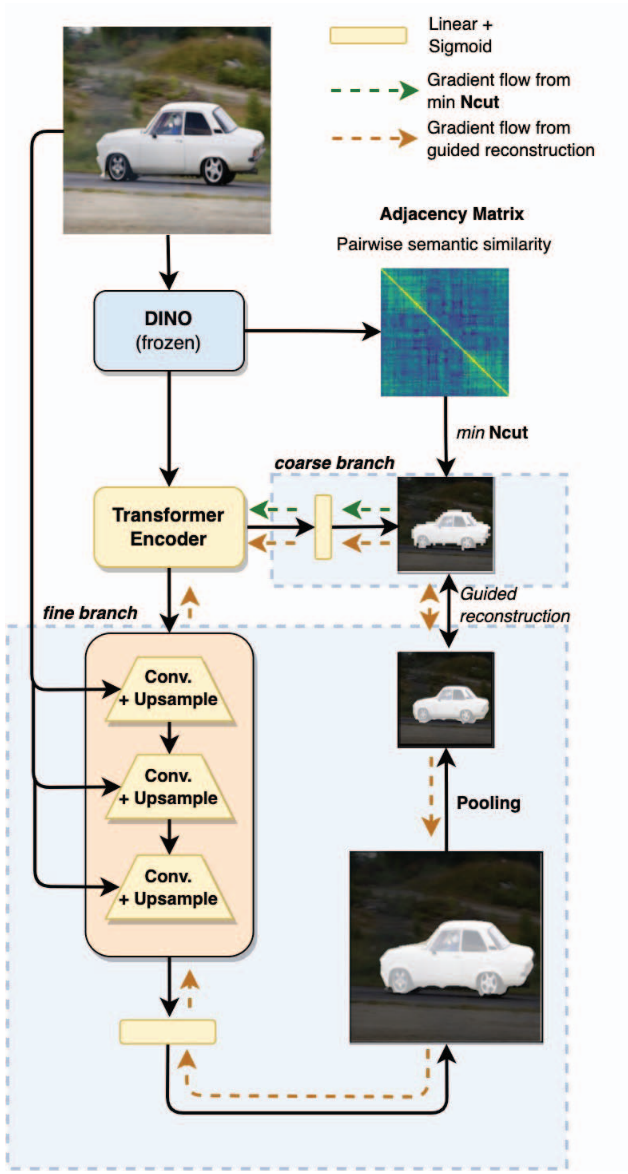

Figure 2 Overview of SEMPART: We refine the SSL features into co-optimized low resolution *coarse* and high resolution *fine masks*, based on graph cut and guided super-resolution respectively.

ject. Distinguishability and predictability of the foreground from the background have also been successful indicators. For example, statistical variations such as in color and texture of the foreground minimally alter the overall distribution of the population [8]. Furthermore, in-painting models such as MAE [22] have been particularly effective at measuring predictability [36] and defining movability [4].

Inferring graph signals [32] for partitioning a semantic graph over an image has gained popularity [39, 54, 41, 30, 1], with recent methods establishing surprisingly strong baselines using traditional techniques. In particular, the solution to the relaxation of the NP-complete discrete normalized cut problem [37] first demonstrated promise in unsupervised image segmentation, which has further translated to recent findings in [39, 54, 40].

[20, 19] discuss the benefit of learning to predict spectral decomposition for a graph and employ graph neural networks in a reinforcement learning setup for predictively performing the normalized cut. More recently, [43] leveraged normalized cut for regularizing a convolutional network driven by partial cross entropy loss in a weakly supervised setting and demonstrated significant performance improvement. More broadly, spectral partitioning of semantic graphs [39, 54, 30, 1] has become an emerging underlying theme for detecting salient regions.

Contributions. In this paper, we propose SEMPART, which builds on ideas from [54, 12, 11] for producing high-quality foreground masks in an SSL setting. SEMPART learns a transformer-based encoder that refines the patchwise DINO features for inferring a relaxation of graph cut that minimizes the expected normalized cut loss [20] over a semantic graph informed by DINO feature correspondences.

As seen in [54, 41, 7], the foreground masks obtained abandon the fine boundary details from processing features at a low resolution. Unlike [39, 54, 41], which perform successive refinement of the *coarse masks* post-inference, SEMPART implements a convolutional *fine branch* that processes and supplements the transformed DINO features with RGB features at progressively increasing resolutions for producing original resolution *fine masks*. Motivated by [12, 11], SEMPART treats the *coarse mask* as the source and the image as a guide for inferring high-quality *fine masks* (see Figure 1, Table 1) regularized by weighted neighborhood-based graph total variation [48].

In summary, our contributions are as follows:

- We propose a novel strategy for co-optimizing *coarse* and *fine masks*, that decouples image partitioning into semantic separation of rich self-supervised features and high-frequency detailing, respectively.
- SEMPART outperforms recent state-of-the-art methods in saliency detection by 3.7% in $\max F_\beta$ and 2.7% in IoU on average and emits high-quality bounding boxes for locating objects.
- SEMPART produces high-quality *fine masks* rapidly by eliminating time-consuming post-inference iterative refinement and saving 200ms on average.

2. Related work

Vision systems have historically benefited from segmenting a scene into objects constituting salient regions [51]. Supervised mechanisms [33, 58] have dominated the landscape despite the prohibitive costs of obtaining labeled data. Traditional unsupervised approaches [5, 29] have encoded beliefs about the foreground region, such as differences in color and contrast and objectness and depth perception, into partitioning techniques.

Spectral methods. Graph-based techniques have received interest wherein spectral partitioning is undertaken over a graphical representation of an image deduced from the priors. [37] proposed *normalized cut* as an improvement over the *min cut* criterion [56], for producing clusters that are well balanced. The relaxation of the discrete problem involved a spectral analysis of the symmetrically normalized graph Laplacian

$$L = I - D^{-1/2} W D^{-1/2}. \tag{1}$$

The unavailability of effective semantic similarity measures between regions of an image for populating the adjacency matrix W inhibited the quality of resulting partitions.

Self-supervised representations. With the emergence of deep techniques for learning contextually aware representations [7, 22, 9, 6], many of these traditional prior-based techniques have demonstrated increased effectiveness and therefore received renewed interest. The semantically aware DINO [7] features were used for implementing seed expansion into salient regions, initialized with patches that are least similar to other patches as seed in LOST [41]. On the contrary, FOUND [42] locates a background seed first and then expands it. In [53], a SOLO [52] model is trained on coarse masks extracted using SSL features, for instance segmentation.

The memory bottleneck of attention mechanism [14] prevents low-resolution deep SSL features from capturing high-frequency details of an image which often are only helpful in predicting coarse masks [41, 39, 54]. Therefore, despite significant performance gains, these methods require computationally heavy post-processing [3, 25, 26] to generate high-quality fine masks.

Inpainting as a helpful object detection tool was first proposed in [36], which hypothesized that it is difficult to predict the foreground given a background and vice versa. SSL features from masked autoencoder (MAE [22]) were also leveraged by recent state-of-the-art MOVE [4] for adversarially training a convolutional mask generator for distinguishing between real- and fake-inpainted images based on movability of salient objects. MOVE established superiority in detecting both salient regions as well as single objects. The movability criterion allows MOVE to directly predict saliency masks at a high resolution which is also why it outperformed its counterparts without post-processing.

SELFMASK [39] uses multi-model SSL features [7, 9, 6] for populating W and constructs pseudo ground truth saliency masks for a subsequent MaskFormer [10] training by clustering eigenvectors of the unnormalized graph Laplacian. Along similar lines, [30] employs clustering based on normalized Laplacian for semantic segmentation and object localization.

Our work is most closely related to [54, 20, 12, 11]. TokenCut [54] makes the bi-partitioning mathematically precise by using the eigenvector with the second smallest eigenvalue, which corresponds to a relaxation of the normalized cut [37] problem and demonstrates value in pursuing graph-based techniques for detecting salient regions.

Iterative computations during inference with expensive post-processing [54, 30], or otherwise training in two stages leveraging multiple SSL models [39] for improving performance, can be limiting. To alleviate this, we follow MOVE's approach of training a single bi-partitioning model as a transformation of the DINO backbone (see Figure 2) and encode our novel strategies into the loss functions (see Section 3.2, Section 3.3). The SEMPART architecture involves a *fine branch* inspired by graph-driven iterative techniques for super-resolution [12, 11] for predicting accurate high-resolution masks.

Minimizing expected graph cut losses over a population was previously evaluated in [20, 19, 1], which proposed to optimize expected normalized cut using graph neural networks. We show that SEMPART exhibits similar benefits (see Table 1, Table 3) from jointly inferred graph-driven bi-partitioning and graph regularized guided super-resolution for generating high-fidelity saliency masks rapidly without any post-processing or multi-stage training.

3. Approach

In this work, we detect salient regions and localize single objects within an image by learning to partition the image into two regions that are semantically less related [54, 21, 39, 1]. We leverage DINO [7], which provides effective pre-trained SSL feature correspondences [54, 21, 30] for learning a *coarse* binary mask that partitions a semantic graph constructed between image patches as nodes. Motivated by image-guided super-resolution [12] and graph regularization [11, 48], we co-optimize and infer masks at the original resolution in parallel, thereby correcting a *coarse mask's* inaccuracies, preserving fine boundary details.

3.1. Background

Normalized Cut. The normalized cut [37] of a weighted undirected complete graph $G = (V, E, w)$ where $w_{ij} > 0$ denotes the weight of $(i, j) \in E$, is given by a binary graph signal $s : v \in V \to s(v) \in \{0, 1\}$ that minimizes

$$\text{Ncut}(A, B) = \frac{w(A, B)}{w(A, V)} + \frac{w(B, A)}{w(B, V)} \tag{2}$$

where $A \coloneqq \{v | v \in V, s(v) = 0\}$, $B \coloneqq \{v | v \in V, s(v) = 1\}$ and $w(A, B) \coloneqq \sum_{s(i)=0, s(j)=1} w_{i,j}$.

Being NP-complete, Shi et al. [37] first proposed to solve a relaxation which amounts to solving a generalized eigensystem followed by discretization. More recently, the relaxation of (4) has been effective at semantically segmenting images in a self-supervised manner [54]. Motivated by [20, 19], non-linear parameterizations of the graph signal have enabled deep partitioning [1] and regularization [43] based on normalized cut.

Deep self-supervised feature correspondences. Large-scale pre-trained self-supervised image embedders such as DINO [7], MAE [22], MoCo [9], SwAV [6] possess beneficial emergent properties for downstream tasks [41, 54, 4, 39, 21]. These models are based on vision transformers [15], which generate an embedding for each patch. Specifically, given an image of dimensions $C \times H \times W$, and an SSL embedder operating with patch size p, we obtain a tensor of size $D \times (H/p \times W/p + 1)$, including the embedding for the `[CLS]` token that represents the entire image. In this paper, we leverage DINO as it emits semantically relevant embeddings [7, 54, 41, 42, 21].

In particular, [54] computed an affinity matrix using the feature correspondences from DINO. A graph view of the output is considered where the graph $G = (V, E)$ contains patches V, and connections between any two patches are encoded in the edge list E. Each patch $v \in V$ has an associated normalized DINO embedding F_v. The affinity matrix is given by the feature correspondences,

$$W_{ij} = \begin{cases} 1 \mid \langle F_{v_i}, F_{v_j} \rangle > \tau \\ \epsilon \mid otherwise. \end{cases} \tag{3}$$

3.2. Self-supervised multi-resolution partitioning (SEMPART)

We propose SEMPART, which converts an image into a semantic graph G over non-overlapping patches, which form the set of nodes V. SEMPART's architecture (see Figure 2) has two main branches that infer a *coarse* and *fine mask* jointly, which are informed by normalized cut and image-guided super-resolution, respectively. We posit that guided super-resolution not only refines the *coarse mask* into a *fine mask* by preserving high-resolution details. It also helps regularize the overall learning and justifies our co-optimization strategy.

Normalized cut for *coarse mask*. A frozen DINO backbone transforms the input image $X \in \mathbb{R}^{3\times320\times320}$ into low-resolution SSL features $F \in \mathbb{R}^{64\times40\times40}$. We apply a single layer transformer encoder with two attention heads, followed by a *coarse branch* (see Figure 2) comprised of a linear classification head, for transforming the low resolution features into a *coarse* saliency mask in the form of a soft partitioning indicator vector $S_{\text{coarse}} \in [0,1]^{|V|}$ where $|V| = 40 \times 40$. For partitions A and B with their indicator vectors $S_A = S_{\text{coarse}}$ and $S_B = 1 - S_A$, (2) is rewritten as

$$\mathcal{L}_{\text{Ncut}}(X) := \text{Ncut}(A, B) = \sum_{i\in\{A,B\}} \frac{S_i^T W(1 - S_i)}{{S_i}^T W \mathbf{1}}. \tag{4}$$

This results in a *coarse mask* at 40×40, which amplifies the semantic distinguishability between the two partitions where the affinity between image patches i and j is computed using the DINO embeddings in (3) and denoted by W_{ij}. Upon minimizing this heuristic over the entire population, we see a significant improvement in performance over solving the generalized eigensystem in [54] (see Table 1).

Guided super-resolution for *fine mask*. The generated *coarse mask* often fails to capture finer high-frequency details [54, 12] at the original image resolution, which is detrimental to the performance in detecting salient regions. Previously, such methods have employed expensive iterative post-processing such as Bilateral Filtering [3, 39, 41, 54] or CRF [25, 21] for every inferenced image. These methods utilize pixels' color and positional information to readjust the generated *coarse masks*. The possibility of erosion of the mask has been discussed as a limitation in [4].

By delegating the generation of linearly separable semantic features to the *coarse branch*, our architecture enables a refinement network to exclusively focus on detailing and denoising at higher frequencies and around the edges. We jointly optimize a *fine branch* (see Figure 2) comprised of a convolutional mask refinement network inspired by a recent guided super-resolution technique [12] which trains a multi-layer perceptron for enhancing the mask with guidance from the image. While [12] performs iterative refinement per image, we co-optimize our refinement network for predicting a *fine mask* which aligns with the *coarse mask* (see Figure 1).

The output from the transformer encoder layer is gradually scaled up from 40×40 to 320×320 in 3 steps. In each step, the image is first scaled up $2\times$ using bilinear interpolation and processed through a convolutional block described in Suppl. Note that we also concatenate the appropriately resized input image to the input of each convolutional block. This information is pertinent for conditioning the *fine branch* to satisfy the regularization in Section 3.3.

The features $\widehat{F} \in \mathbb{R}^{131\times320\times320}$ from the last convolutional block are linearly classified into $S_{\text{fine}} \in [0,1]^{320\times320}$ which is subsequently average pooled to $\widehat{S}_{\text{fine}} \in [0,1]^{40\times40}$ for aligning with the $S_{\text{coarse}} \in [0,1]^{40\times40}$. The corresponding loss function is given as

$$\mathcal{L}_{\text{SR}}(X) := \|\widehat{S}_{\text{fine}} - S_{\text{coarse}}\|_2^2. \tag{5}$$

3.3. Graph total variation regularization (GTV)

Graph-based regularization has yielded benefits in capturing high-frequency details of an image in [11, 12]. A

	Model	DUT-OMRON [57]			DUTS-TE [49]			ECSSD [38]		
		Acc	**IoU**	**max**F_β	**Acc**	**IoU**	**max**F_β	**Acc**	**IoU**	**max**F_β
Method	LOST [41]	.797	.410	.473	.871	.518	.611	.895	.654	.758
	TokenCut [54]	.880	.533	.600	.903	.576	.672	.918	.712	.803
	FreeSOLO [53]	.909	.560	.684	.924	.613	.750	.917	.703	.858
	MOVE [4]	.923	.615	.712	.950	.713	.815	.954	.830	.916
	SEMPART-Coarse	**.932**	.640	.755	.956	.727	.864	.961	.837	.943
	SEMPART-Fine	**.932**	**.668**	**.764**	**<u>.959</u>**	**<u>.749</u>**	**.867**	**.964**	**.855**	**<u>.947</u>**
+ BF	LOST+BF	.818	.489	.578	.887	.572	.697	.916	.723	.837
	TokenCut+BF	.897	.618	.697	.914	.624	.755	.934	.772	.874
	MOVE+BF	.931	.636	.734	.951	.687	.821	.953	.801	.916
	SEMPART-Coarse+BF	**.934**	**.661**	**.764**	**.957**	**.697**	**.858**	**.960**	**.820**	**.932**
	SEMPART-Fine+BF	.933	.653	.760	.955	.685	.853	.959	.816	.931
+ SELFMASK	SELFMASK on pseudo + BF [39]	.919	.655	(.774)*	.933	.660	(.819)*	.955	.818	(.911)*
	SELFMASK on MOVE	.933	.666	.756	.954	.728	.829	.956	.835	.921
	SELFMASK on MOVE + BF	.937	.665	.766	.952	.687	.827	.952	.800	.917
	SELFMASK on SEMPART-Coarse	.936	.675	.773	**.958**	.743	.872	.962	.843	.938
	SELFMASK on SEMPART-Fine	**<u>.942</u>**	**<u>.698</u>**	**<u>.799</u>**	**.958**	**<u>.749</u>**	**<u>.879</u>**	**.963**	**.850**	**.944**

* The $\max F_\beta$ for SELFMASK on pseudo + BF is reported within brackets (), from the reevaluation in [4] upon confirming that it was originally calculated incorrectly [4, 42].

Table 1 Quantitative comparison of SEMPART with state-of-the-art MOVE and other related works for saliency detection. SEMPART-Coarse and -Fine outperform MOVE significantly in all three evaluation categories (Method, +BF, +SELFMASK) across all datasets. The best-performing method in a category and across categories is in **bold** and **<u>underlined</u>**, respectively.

similarity metric between pixels of an image X is used to populate the affinity matrix $A > 0$, which is then used to compute the degree matrix D. The graph Laplacian $L = D - A$ is used to compute the graph regularizer as the quadratic form for a graph signal [32] s, given by

$$\mathcal{L}_{reg} = \frac{1}{2} \sum_{(i,j)\in E} A_{ij}(s(i) - s(j))^2. \tag{6}$$

Considering significant computational complexity from the total number of pairs of pixels, we enforce $A_{ij} = 0$ when pixels X_i and X_j are not vertically or horizontally adjacent, also known as the pixel neighborhood $\mathcal{N}$. This is equivalent to a weighted version of the total variation (TV) loss [28, 16], which has been previously used for denoising images and other signals [2, 23, 16, 34]. A natural extension to graphs is discussed in [48].

GTV fine. The guided super-resolution can result in more than one *fine mask* for a given *coarse mask*, which is where our graph total variation (GTV) loss not only works as a denoiser but plays a more important role as a regularizer. More specifically, $A_{ij} = \exp\left(-\|X_i - X_j\|_2^2/\sigma\right)$ is given by the euclidean similarity between the pairwise pixels. As a result, the $\mathcal{L}_{\text{GTV-fine}}$ loss encourages the upsampler along the *fine* branch in Figure 2 to leverage the color information.

GTV coarse. We also implement a similar graph TV regularizer denoted by $\mathcal{L}_{\text{GTV-coarse}}$ for the *coarse mask* based on $A_{ij} = W_{ij}\mathbf{1}\{i \in \mathcal{N}(j)\}$ where W_{ij} is as defined in (3). This is responsible for denoising and predicting a smooth *coarse mask*.

3.4. Loss formulation

The SEMPART losses in Section 3.2 together with the GTV losses in Section 3.3 drive the joint learning of *coarse* and *fine masks*. While the SEMPART losses are driven by DINO feature correspondences for inferring accurate image partitions, the GTV losses are significantly involved in denoising the predicted masks and regularizing the overall learning process. The loss functions for the *coarse* and *fine branches*, respectively, are,

$$\begin{aligned} \mathcal{L}_{\text{coarse}}(x) &= \mathcal{L}_{\text{Ncut}}(x) + \lambda_{\text{GTV-coarse}}\mathcal{L}_{\text{GTV-coarse}}(x) \\ \mathcal{L}_{\text{fine}}(x) &= \lambda_{\text{GTV-fine}}\mathcal{L}_{\text{GTV-fine}}(x) \\ \mathcal{L}_{\text{joint}}(x) &= \lambda_{\text{SR}}\mathcal{L}_{\text{SR}}(x). \end{aligned} \tag{7}$$

This gives us our final expected self-supervised loss function $\mathcal{L}_{\text{SEMPART}} = \underset{x\sim\mathbb{P}(X)}{\mathbb{E}}[\mathcal{L}_{\text{coarse}}(x) + \mathcal{L}_{\text{fine}}(x) + \mathcal{L}_{\text{joint}}(x)]$.

4. Experiments

As done in [54, 4], we evaluate SEMPART on unsupervised saliency segmentation and single object detection.

4.1. Implementation

In our work, we use the self-supervised [7] ViT-s/8 transformer from the official implementation of DINO [7]. DINO uses an 8×8 non-overlapping patch on a $3\times320\times320$

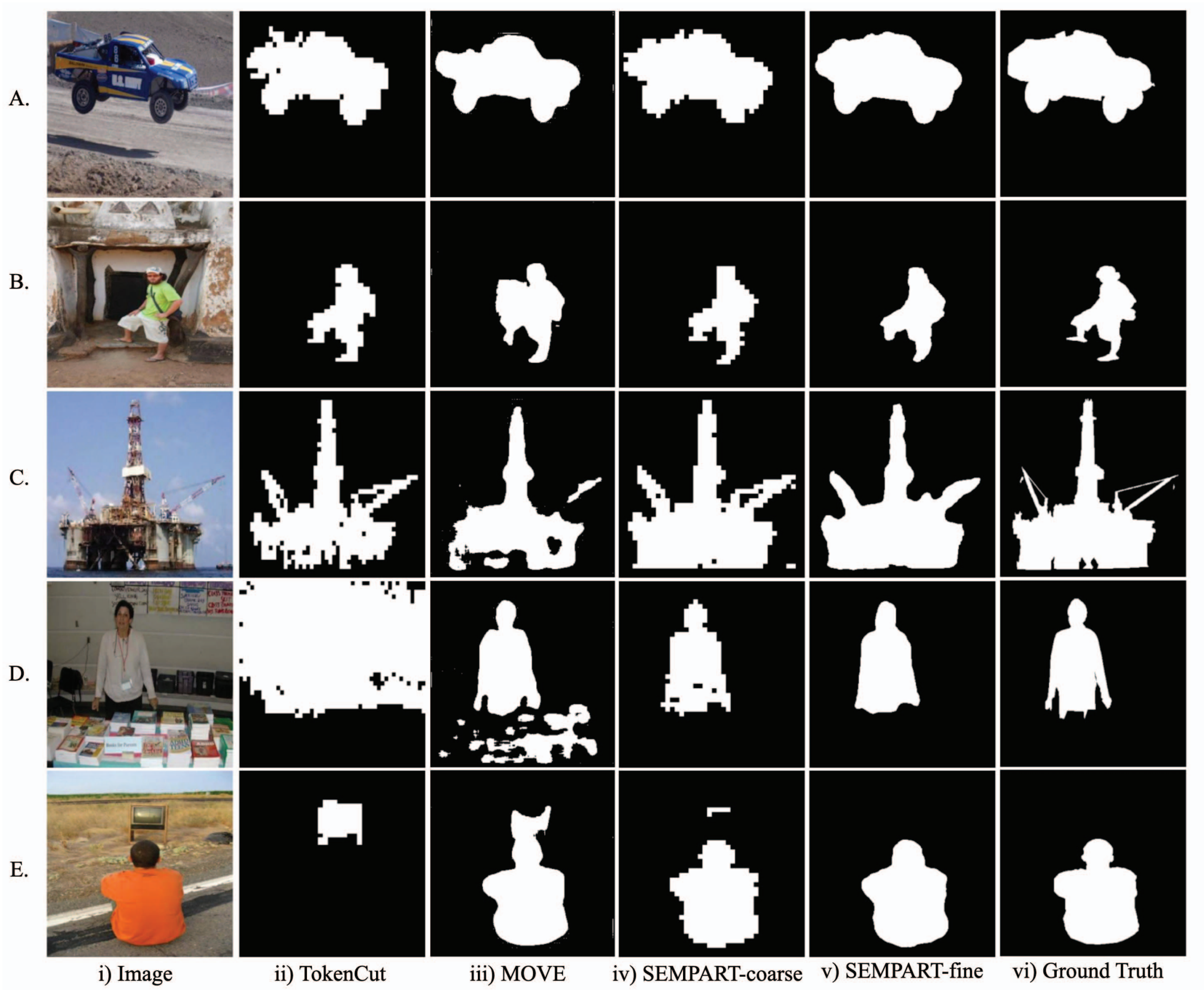

Figure 3 Qualitative comparison of SEMPART-coarse and -fine with TokenCut [54] and MOVE [4] for samples from DUT-OMRON [57].

input and emits $384 \times 40 \times 40$ output which is provided to our simple transformer encoder layer and then routed through both the *coarse* and *fine branches* in Figure 2. We employ Adam optimizer [24] with a learning rate of 0.0001 and $\beta = (0.9, 0.999)$. We implemented SEMPART in PyTorch and trained our models for 20 epochs with a batch size of 8 on a single NVIDIA Tesla P40 GPU. Following careful consideration, hyperparameters $\lambda_{\text{GTV-coarse}} = 0.0006$, $\lambda_{\text{SR}} = 20$, $\lambda_{\text{GTV-fine}} = 0.0002$ have been applied for all results of SEMPART.

Graph affinity. (a) Normalized cut. Our implementation follows [54] in computing the affinity matrix W based on (3) with a minor deviation. We set $W_{ii} = 0$ to discard self-loops that do not belong to a graph cut. We show empirically that this improves model performance. Additionally we set $\tau = 0.2$ and ϵ =1e-6 in (3) for the $\mathcal{L}_{\text{Ncut}}$ loss. **(b) GTV Coarse.** In addition to details provided in Section 3.3, we set $\tau = 0$ and ϵ =1e-6 for numerical stability. **(c) GTV Fine.** $\mathcal{L}_{\text{GTV-fine}}$ regularizes the *fine mask* by limiting the possible solutions. The convolutional blocks learn to generate features that leverage both the contextual features from the transformer encoder and the RGB image features for predicting fine masks that mimic the *coarse mask* but also preserve the high-frequency image details. In addition to details provided in Section 3.3, we also set $\sigma = 1$.

Foreground selection. We first binarize the indicator vector with threshold = 0.5. In order to pick the foreground, we consider four strategies. **(a)** Select the patch with a lower average distance to the center as the foreground. **(b)** Discarding partitions with full spatial width or height as background, selecting the smaller partition to break a tie. **(c)** Select the partition with greatest attention from the last layer of DINO. **(d)** Select the partition occupying the least number of corners. If there is a tie, select the smaller partition.

4.2. Unsupervised saliency segmentation

Datasets. As done in [4, 1, 39], we trained SEMPART on the train split of DUTS [49], known as DUTS-TR and evaluate the performance of our model on the corresponding test split DUTS-TE [49], as well as DUT-OMRON [57] and

Method	Avg. Time	Model	RES	GPU	CPU
TokenCut	130ms	No	Low	Yes	Yes
TokenCut+BF	337ms	No	High	Yes	Yes
MOVE	13ms	Yes	High	Yes	No
SEMPART	14ms	Yes	High	Yes	No

Table 2 Both SEMPART and MOVE train a model, generate high-resolution masks, and have comparable average inference times per image.

ECSSD [38]. DUTS-TR contains 10,553 images, DUTS-TE contains $5,019$ images, DUT-OMRON contains 5,168 images, and ECSSD contains 1000 images.

Evaluation. As done in [4, 54], we compute the per-pixel mask accuracy (Acc), intersection over union (IoU), and $\max F_\beta$ [54] for evaluating the performance of SEMPART. Accuracy is the fraction of pixels correctly predicted into the foreground or background. The overlap between the binary saliency mask and the ground truth gives IoU. We set $\beta = 0.3$ as per [4, 54] where $\max F_\beta$ is given for the threshold used for binarizing the mask that maximizes F_β.

Results. We compared the performance of SEMPART with recent state-of-the-art MOVE [4] and several other standard baselines referenced therein. Table 1 contains three horizontal sections corresponding to the baseline method, followed by applying a bilateral filtering [3] step. The final section involves generating pseudo ground truth based on the baseline method and training MaskFormer [10] in a class agnostic manner, as in [39].

We observe that applying the bilateral filter after inferencing SEMPART on a per-image basis is detrimental to the overall performance, as is also seen in [4], with the performance of SEMPART-Fine deteriorating significantly.

SEMPART significantly outperformed all other baselines in all three sections across all datasets. Although SEMPART is primarily motivated by the normalized cut minimization in [54], the expected normalized cut loss in Section 3.4 co-optimized with the image-guided graph-based super-resolution loss results in significant improvement in performance. As seen in Figure 3, the per-image optimization in TokenCut selects regions that are not salient or present in the foreground.

SEMPART significantly outperforms the movability [4] heuristic in all three sections for all datasets. From Figure 3, we find that MOVE may include multiple semantically unrelated patches into the movable object mask. Additionally, we note that MOVE greatly relies on retraining according to SELFMASK [39] for outperforming previous state-of-the-art. While SEMPART-Coarse predicts noisy masks (see Figure 3-A,C,D, and E) with slight errors as seen in the last example, SEMPART-Fine results in refinement with improved ground truth alignment.

Figure 4 SEMPART for single object detection. Green boxes are ground truth bounding boxes, and the red box is our predicted bounding box. Intersection area is highlighted.

4.3. Single object detection

Datasets. We evaluate our model on three datasets - the train split of COCO20K [27] and the training and validation splits of VOC07 [17] and VOC12 [18]. Each image in these datasets has one or more bounding boxes corresponding to each object. The objective is to localize any single object.

Evaluation. We detect connected components for separating multiple objects for an image's SEMPART mask[1]. The component with the largest bounding box is used as the object prediction. Suppose the highest IoU between our predicted bounding box and all ground truth bounding boxes exceeds 0.5. In that case, we treat it as a successful prediction and use this to compute *Correct Localization* (CorLoc) metric which is simply the accuracy of prediction.

Results. SEMPART results in superior bounding-boxes which perform comparably with state-of-the-art MOVE, outperforming it on COCO20k dataset (see Table 3). Our findings suggest that increasing τ to 0.25 helps us prevent co-located disparate objects from lying in the same connected component and results in a slight improvement.

4.4. Ablations

We ablated SEMPART for saliency segmentation as follows,

Foreground selection. Unlike [4], where the foreground is given by the *movable* object, SEMPART selects partitions based on occupying *least corners* given by SEMPART-Fine

[1] If multiple objects lie in a component this evaluation is less reliable.

Method	VOC07	VOC12	COCO20K
DDT+ [55]	50.2	53.1	38.2
rOSD [44]	54.5	55.3	48.5
LOD [45]	53.6	55.1	48.5
FreeSOLO [53]	56.1	56.7	52.8
LOST [41]	61.9	64.0	50.7
Deep Spectral [30]	62.7	66.4	52.2
TokenCut [54]	68.8	72.1	58.8
MOVE [4]	**76.0**	**78.8**	66.6
SEMPART-Coarse	74.7	77.4	**66.9**
SEMPART-Fine	75.1	76.8	66.4

Table 3 SEMPART bounding boxes exhibits a high CorLoc comparable to state-of-the-art MOVE [4], for single object discovery on VOC2007 [17], VOC2012[18] and outperforms it on COCO20K [27] dataset.

Method	OMRON*	D-TE*	ECSSD
fs: framing prior	0.663	0.730	0.825
fs: centrality	0.652	0.736	0.854
fs: total attention	**0.668**	0.745	0.853
w/ self-loops in W	0.667	0.743	0.846
w/o GTV coarse	0.646	**0.749**	0.848
w/o GTV fine	0.637	0.717	0.818
train fine mask directly	0.645	0.738	0.845
w/o joint training	0.662	0.743	0.849
SEMPART-Fine	**0.668**	**0.749**	**0.855**

Table 4 Ablations of SEMPART for saliency, using mIoU. *Shorthand has been used due to space constraints; OMRON refers to DUT-OMRON [57] and D-TE refers to DUTS-TE [49]; *fs* denotes foreground selection.

in Table 4. Motivated by [39, 41], we compare with selection based on closeness to the image center (*centrality*), as well as the *framing prior* [39], which labels the segment occupying full spatial width or height as background while breaking ties based on selecting the smaller partition as foreground. Another heuristic that is a close contender to *least corners* is *total attention*, in which the partition having the highest total overlap with the DINO `[CLS]` token attention map as foreground.

Self-loops. We populate W_{ii} with (3) instead of 0 for $\mathcal{L}_{\text{Ncut}}$ and demonstrate that the performance deteriorates.

Graph TV regularization. SEMPART without either $\mathcal{L}_{\text{GTV-coarse}}$ or $\mathcal{L}_{\text{GTV-fine}}$ is detrimental to performance. The absence of the GTV-fine loss has a greater negative impact.

Training fine mask directly. We evaluate a setting where we only have a *fine branch* (see Figure 2), and the $\mathcal{L}_{\text{Ncut}}$ and $\mathcal{L}_{\text{GTV-fine}}$ losses. Table 4 demonstrates that this is inferior to SEMPART despite being almost equivalent in the number of parameters. We attribute this to the absence of the *coarse branch* and the corresponding $\mathcal{L}_{\text{Ncut}}$ loss which in turn regularized the transformer encoder for subsequent consumption by the convolutional blocks.

Joint training. We evaluate a variant of SEMPART, where the *coarse* and *fine branch* are trained independently. While the $\mathcal{L}_{\text{coarse}}$ only optimizes the *coarse branch* and the transformer encoder (see deviations from Figure 2 in Suppl.), the gradients from $\mathcal{L}_{\text{fine}}$ and $\mathcal{L}_{\text{joint}}$ are prohibited from optimizing these modules. As seen in Table 4, this is detrimental to performance on all datasets, verifying our hypothesis that co-optimizing *coarse* and *fine* mask is mutually beneficial.

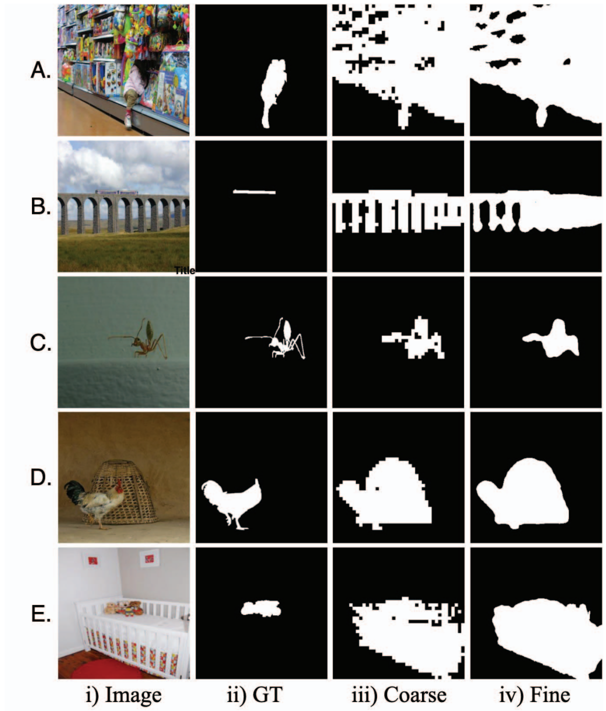

Figure 5 Limitations of SEMPART. Human bias towards humans and moving objects are shown in A and B. SEMPART cannot capture the intricate details and smooths over narrow regions in B and C. An immovable background object is included, which is not as visually salient as the rooster in D. The crib is the same color as the wall in E; therefore, the toys are prominent. However, DINO highlights the semantic differences for partitioning the entire crib from the background.

4.5. Limitations

Visual saliency is not agnostic to various human biases in favor of humans and animals, as well as objects which are likely to move in a subsequent frame or have high contrast with the background. Figure 5 discusses examples where the ground truth favors a human, a train crossing a bridge, and a rooster over all other objects. SEMPART results in over-selection here as it does not explicitly incorporate these priors or even control the object size. Furthermore, our graph TV loss can sometimes merge narrow co-located regions into the mask, as seen in the Figure 5-B and C, which can also be detrimental to localizing objects.

5. Conclusion

SEMPART demonstrates the efficacy of graph-driven objectives towards self-supervised image partitioning and establishes state-of-the-art performance for detecting salient regions and a competitive advantage in localizing objects. We address the limitations of expensive post-processing, limited resolution, and noise artifacts in saliency masks. We demonstrate the value of a joint learning paradigm for inferring high-quality masks at multiple resolutions using SEMPART, which will hopefully be a vital enabler of the subsequent investigation into class-aware object detection for diverse vision systems.

Acknowledgements The authors gratefully thank Ambareesh Revanur and Deepak Pai for their valuable feedback, and the anonymous reviewers for their comments.

References

[1] Amit Aflalo, Shai Bagon, Tamar Kashti, and Yonina C. Eldar. Deepcut: Unsupervised segmentation using graph neural networks clustering. *CoRR*, abs/2212.05853, 2022. 2, 3, 4, 6

[2] William K. Allard. Total variation regularization for image denoising, i. geometric theory. *SIAM Journal on Mathematical Analysis*, 39(4):1150–1190, 2008. 5

[3] Jonathan T. Barron and Ben Poole. The fast bilateral solver. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III*, volume 9907 of *Lecture Notes in Computer Science*, pages 617–632. Springer, 2016. 3, 4, 7

[4] Adam Bielski and Paolo Favaro. MOVE: unsupervised movable object segmentation and detection. *CoRR*, abs/2210.07920, 2022. 1, 2, 3, 4, 5, 6, 7, 8, 12, 15

[5] Ali Borji, Ming-Ming Cheng, Huaizu Jiang, and Jia Li. Salient object detection: A survey. *CoRR*, abs/1411.5878, 2014. 1, 2

[6] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 1, 3, 4

[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 9630–9640. IEEE, 2021. 1, 2, 3, 4, 5, 12, 19

[8] Mickaël Chen, Thierry Artières, and Ludovic Denoyer. Unsupervised object segmentation by redrawing. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 12705–12716, 2019. 2

[9] Xinlei Chen, Haoqi Fan, Ross B. Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *CoRR*, abs/2003.04297, 2020. 1, 3, 4

[10] Bowen Cheng, Alexander G. Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 17864–17875, 2021. 3, 7

[11] Riccardo de Lutio, Alexander Becker, Stefano D'Aronco, Stefania Russo, Jan D. Wegner, and Konrad Schindler. Learning graph regularisation for guided super-resolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 1969–1978. IEEE, 2022. 2, 3, 4

[12] Riccardo de Lutio, Stefano D'Aronco, Jan Dirk Wegner, and Konrad Schindler. Guided super-resolution as pixel-to-pixel transformation. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 8828–8836. IEEE, 2019. 2, 3, 4

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 19

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. 3

[15] 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 worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 4, 12

[16] Vania Vieira Estrela, Hermes Aguiar Magalhaes, and Osamu Saotome. Total variation applications in computer vision. *CoRR*, abs/1603.09599, 2016. 5

[17] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html. 7, 8, 19

[18] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html. 7, 8, 19

[19] Alice Gatti, Zhixiong Hu, Tess Smidt, Esmond G. Ng, and Pieter Ghysels. Graph partitioning and sparse matrix ordering using reinforcement learning and graph neural networks. *Journal of Machine Learning Research*, 23(303):1–28, 2022. 2, 3, 4

[20] Alice Gatti, Zhixiong Hu, Tess E. Smidt, Esmond G. Ng, and Pieter Ghysels. Deep learning and spectral embedding for graph partitioning. In Xiaoye S. Li and Keita Teranishi, editors, *Proceedings of the 2022 SIAM Conference on Parallel Processing for Scientific Computing, PPSC 2022, Seattle, WA, USA, February 23-26, 2022*, pages 25–36. SIAM, 2022. 2, 3, 4

[21] Mark Hamilton, Zhoutong Zhang, Bharath Hariharan, Noah Snavely, and William T. Freeman. Unsupervised semantic segmentation by distilling feature correspondences. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. 1, 3, 4

[22] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. Masked autoencoders are scalable vision learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 15979–15988. IEEE, 2022. 1, 2, 3, 4

[23] Xianxu Hou, Linlin Shen, Or Patashnik, Daniel Cohen-Or, and Hui Huang. Feat: Face editing with attention, 2022. 5

[24] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 6

[25] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In John Shawe-Taylor, Richard S. Zemel, Peter L. Bartlett, Fernando C. N. Pereira, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain*, pages 109–117, 2011. 3, 4

[26] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Carla E. Brodley and Andrea Pohoreckyj Danyluk, editors, *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001), Williams College, Williamstown, MA, USA, June 28 - July 1, 2001*, pages 282–289. Morgan Kaufmann, 2001. 3

[27] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Doll'a r, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014. 7, 8, 19

[28] Jiaming Liu, Yu Sun, Xiaojian Xu, and Ulugbek S. Kamilov. Image restoration using total variation regularized deep image prior. *CoRR*, abs/1810.12864, 2018. 5

[29] Tie Liu, Jian Sun, Nanning Zheng, Xiaoou Tang, and Heung-Yeung Shum. Learning to detect A salient object. In *2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2007), 18-23 June 2007, Minneapolis, Minnesota, USA*. IEEE Computer Society, 2007. 2

[30] Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. Deep spectral methods: A surprisingly strong baseline for unsupervised semantic segmentation and localization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 8354–8365. IEEE, 2022. 1, 2, 3, 8

[31] Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. Finding an unsupervised image segmenter in each of your deep generative models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. 1

[32] Antonio Ortega, Pascal Frossard, Jelena Kovačević, José M. F. Moura, and Pierre Vandergheynst. Graph signal processing: Overview, challenges, and applications. *Proceedings of the IEEE*, 106(5):808–828, 2018. 2, 5

[33] Xuebin Qin, Zichen Zhang, Chenyang Huang, Masood Dehghan, Osmar R. Zaïane, and Martin Jägersand. U^2-net: Going deeper with nested u-structure for salient object detection. *Pattern Recognit.*, 106:107404, 2020. 2, 14, 15

[34] Ambareesh Revanur, Debraj Basu, Shradha Agrawal, Dhwanit Agarwal, and Deepak Pai. Coralstyleclip: Co-optimized region and layer selection for image editing, 2023. 5

[35] Denise Rey and Markus Neuhäuser. Wilcoxon-signed-rank test. In *International Encyclopedia of Statistical Science*, 2011. 14

[36] Pedro Savarese, Sunnie S. Y. Kim, Michael Maire, Greg Shakhnarovich, and David McAllester. Information-theoretic segmentation by inpainting error maximization. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 4029–4039. Computer Vision Foundation / IEEE, 2021. 1, 2, 3

[37] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):888–905, 2000. 2, 3

[38] Jianping Shi, Qiong Yan, Li Xu, and Jiaya Jia. Hierarchical image saliency detection on extended CSSD. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(4):717–729, 2016. 5, 7, 17, 19

[39] Gyungin Shin, Samuel Albanie, and Weidi Xie. Unsupervised salient object detection with spectral cluster voting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2022, New Orleans, LA, USA, June 19-20, 2022*, pages 3970–3979. IEEE, 2022. 1, 2, 3, 4, 5, 6, 7, 8

[40] Gyungin Shin, Weidi Xie, and Samuel Albanie. Namedmask: Distilling segmenters from complementary foundation models. *CoRR*, abs/2209.11228, 2022. 2, 19

[41] 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 *32nd British Machine Vision Conference 2021, BMVC 2021, Online, November 22-25, 2021*, page 310. BMVA Press, 2021. 1, 2, 3, 4, 5, 8

[42] Oriane Siméoni, Chloé Sekkat, Gilles Puy, Antonín Vobecký, Éloi Zablocki, and Patrick Pérez. Unsupervised object localization: Observing the background to discover objects. *CoRR*, abs/2212.07834, 2022. 1, 3, 4, 5, 19

[43] Meng Tang, Abdelaziz Djelouah, Federico Perazzi, Yuri Boykov, and Christopher Schroers. Normalized cut loss for weakly-supervised CNN segmentation. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 1818–1827. Computer Vision Foundation / IEEE Computer Society, 2018. 2, 4

[44] Huy V. Vo, Patrick Pérez, and Jean Ponce. Toward unsupervised, multi-object discovery in large-scale image collections. *CoRR*, abs/2007.02662, 2020. 8

[45] Van Huy Vo, Elena Sizikova, Cordelia Schmid, Patrick Pérez, and Jean Ponce. Large-scale unsupervised object discovery. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 16764–16778. Curran Associates, Inc., 2021. 8

[46] Andrey Voynov and Artem Babenko. Unsupervised discovery of interpretable directions in the GAN latent space. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 9786–9796. PMLR, 2020. 1

[47] Andrey Voynov, Stanislav Morozov, and Artem Babenko. Object segmentation without labels with large-scale generative models. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 10596–10606. PMLR, 2021. 1

[48] Huy Vu, Gene Cheung, and Yonina C. Eldar. Unrolling of deep graph total variation for image denoising. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Toronto, ON, Canada, June 6-11, 2021*, pages 2050–2054. IEEE, 2021. 2, 3, 5

[49] 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*, 2017. 5, 6, 8, 18, 19

[50] Peng Wang, Jingdong Wang, Gang Zeng, Jie Feng, Hongbin Zha, and Shipeng Li. Salient object detection for searched web images via global saliency. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3194–3201, 2012. 19

[51] Wenguan Wang, Qiuxia Lai, Huazhu Fu, Jianbing Shen, and Haibin Ling. Salient object detection in the deep learning era: An in-depth survey. *CoRR*, abs/1904.09146, 2019. 1, 2

[52] Xinlong Wang, Tao Kong, Chunhua Shen, Yuning Jiang, and Lei Li. SOLO: segmenting objects by locations. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XVIII*, volume 12363 of *Lecture Notes in Computer Science*, pages 649–665. Springer, 2020. 3

[53] Xinlong Wang, Zhiding Yu, Shalini De Mello, Jan Kautz, Anima Anandkumar, Chunhua Shen, and Jose M. Alvarez. Freesolo: Learning to segment objects without annotations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 14156–14166. IEEE, 2022. 3, 5, 8

[54] 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. *arXiv preprint arXiv:2209.00383*, 2022. 1, 2, 3, 4, 5, 6, 7, 8

[55] Xiu-Shen Wei, Chen-Lin Zhang, Jianxin Wu, Chunhua Shen, and Zhi-Hua Zhou. Unsupervised object discovery and co-localization by deep descriptor transforming, 2017. 8

[56] Zhenyu Wu and Richard M. Leahy. An optimal graph theoretic approach to data clustering: Theory and its application to image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 15(11):1101–1113, 1993. 3

[57] Chuan Yang, Lihe Zhang, Ruan Xiang Lu, Huchuan, and Ming-Hsuan Yang. Saliency detection via graph-based manifold ranking. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3166–3173. IEEE, 2013. 5, 6, 8, 16, 19

[58] Yi Ke Yun and Weisi Lin. Selfreformer: Self-refined network with transformer for salient object detection. *CoRR*, abs/2205.11283, 2022. 2, 14, 15

[59] Yuan Zhou, Ailing Mao, Shuwei Huo, Jianjun Lei, and Sun-Yuan Kung. Salient object detection via fuzzy theory and object-level enhancement. *IEEE Transactions on Multimedia*, 21(1):74–85, 2019. 1