A. Notation

For an image $X \in \mathbb{R}^{3\times320\times320}$, we represent the self-supervised features of DINO [7] obtained for all 8×8 non overlapping patches as $F \in \mathbb{R}^{384\times40\times40}$. We use $\mathcal{L}$ to denote loss functions. $\mathbb{P}(M)$ and $\mathbb{E}[M]$ denote the distribution and the expected value of the random variable M. $\mathbf{1}\{\cdot\}$ is used to denote the indicator function.

For a graph, $G = (V, E)$, V, and E denote the vertex and edge set, respectively. W and A represent the adjacency or affinity matrix for the $\mathcal{L}_{\text{Ncut}}$ in Section 3.2 and the GTV losses in Section 3.3 respectively.

I denotes the identity matrix. D and L correspond to the degree matrix and the Laplacian matrix for the graph G, respectively. $s : v \in V \to s(v) \in R$ has been used to denote a scalar signal as a function defined over the graph's nodes $v \in V$ as the domain. The definition of S naturally follows as $S := [s(1), s(2), \ldots, s(|V|)]^T$.

B. Architecture for Section 3

Section 3 describes the essential details of the SEMPART architecture in which we emphasize the importance of two vital learnable components: **(a)** the transformer encoder as a shared parametrized module between both the *coarse* and *fine branch*, **(b)** the convolutional mask refinement network for generating high resolution *fine masks*. Figure 6 (a) and (b) presents the transformer encoder as well as the convolutional mask refinement network, respectively, in greater detail. Furthermore, we also elaborate upon these individual modules in Appendix C.

C. Pseudocode for Section 3

SEMPART is a self-supervised multi-resolution image bi-partitioning heuristic that successfully distills the encoded information from DINO [7] towards high-quality unsupervised semantically meaningful partitions that significantly resonate with the notion of visual saliency for an image. In this section, we elaborate upon the forward pass described in Section 3.2 to Section 3.4 culminating in Algorithm 1.

DINO backbone [7]: DINO [7] is a widely adopted self-supervised vision model which emits features that are contextually aware and captures the semantic richness of an image (see [7, Figure 1]). SEMPART leverages the self-supervised [7] ViT-s/8 transformer based on [15] from the official implementation of DINO [7], which processes a 320×320 image X as a 40×40 positionally aware flattened sequence of 8×8 non overlapping patches. We denote the transformation by

$$\text{DINO}(X) : X \in \mathbb{R}^{3\times320\times320} \to F \in \mathbb{R}^{384\times40\times40}. \quad (8)$$

Note that in fact DINO emits $\mathbb{R}^{384\times(1+40\times40)}$, however we discard the `[CLS]` token feature for subsequent modules. In our implementation, the DINO backbone remains frozen.

Transformer encoder [15]: We apply a single layer transformer encoder[2] with two attention heads that transform $F \in \mathbb{R}^{384\times40\times40}$ to $\widetilde{F} \in \mathbb{R}^{64\times40\times40}$.

$$\widetilde{F} \leftarrow \textsc{TransformerEncoder}(F). \quad (9)$$

Emitted features $\widetilde{F}$ are shared between both the SEMPART-Coarse and SEMPART-Fine branches (see Figure 2).

Convolutional mask refinement network (Section 3.2): As also done in [4], we define $\textsc{Block}^{\text{in\_ch}}_{\text{out\_ch}}$ as

$$3 \times 3\ \textsc{Conv}^{\text{in\_ch}}_{\text{out\_ch}} \to \textsc{BatchNorm} \to \textsc{LeakyReLU} \quad (10)$$

where $K \times K\ \textsc{Conv}^{\text{in\_ch}}_{\text{out\_ch}}$ is a padded $K \times K$ convolution with stride = 1, in_ch and out_ch correspond to the number of input and output channels respectively. Before each block, we also concatenate - denoted by the $||_c$ operator - an appropriately resized image along the channel dimension.

Consequently, our convolutional mask refinement network is given by alternating bilinear UPSAMPLE and BLOCK as follows

$$\begin{aligned}
\widetilde{F}' &\leftarrow \textsc{Block}^{67}_{192}\left[\textsc{Upsample}^{2\times2}_{\text{bilinear}}\left(\widetilde{F}\right) ||_c X^{3\times80\times80}\right] \\
\widetilde{F}'' &\leftarrow \textsc{Block}^{195}_{128}\left[\textsc{Upsample}^{2\times2}_{\text{bilinear}}\left(\widetilde{F}'\right) ||_c X^{3\times160\times160}\right] \\
\widetilde{F}''' &\leftarrow \textsc{Block}^{131}_{128}\left[\textsc{Upsample}^{2\times2}_{\text{bilinear}}\left(\widetilde{F}''\right) ||_c X^{3\times320\times320}\right] \\
\widehat{F} &\leftarrow \textsc{Block}^{128}_{128}\left(\widetilde{F}'''\right) ||_c X^{3\times320\times320}. \quad (11)
\end{aligned}$$

The image X is provided as side information and is essential for conditioning the convolutional mask refinement network towards generating *fine masks* driven by the $\mathcal{L}_{\text{GTV-fine}}$ loss. We modularize the complete convolutional mask refinement transformation given in (11) as follows,

$$\widehat{F} \leftarrow \textsc{ConvMaskRefine}(\widetilde{F}, X). \quad (12)$$

Coarse branch (Section 3.2): The *coarse branch* applies a binary linear classification head (LCH) as a composition of a linear layer followed by sigmoid to $\widetilde{F}$, resulting in $S_{\text{coarse}} \in [0, 1]^{40\times40}$.

$$S_{\text{coarse}} \leftarrow \text{LCH}^{64}_{1}\left(\widetilde{F}\right). \quad (13)$$

Here $\text{LCH}^{\text{in\_ch}}_{1}$ corresponds to

$$1 \times 1\ \textsc{Conv}^{\text{in\_ch}}_{1} \to \textsc{sigmoid}. \quad (14)$$

We denote this operation as follows

$$S_{\text{coarse}} \leftarrow \textsc{CoarseBranch}(\widetilde{F}). \quad (15)$$

[2] Implementation is borrowed from [15].

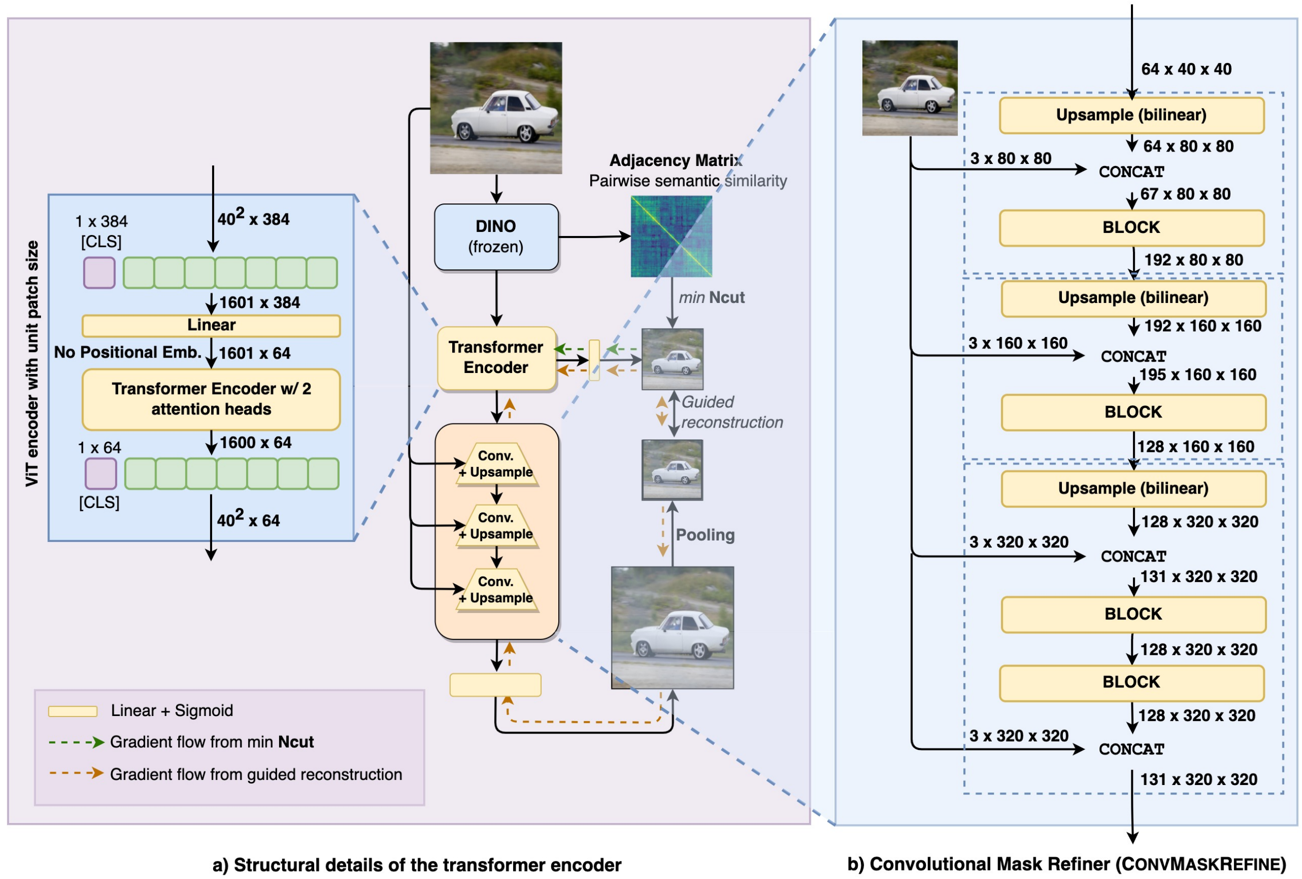

Figure 6 Expanded overview of SEMPART: In addition to the details presented in Figure 2, we zoom in to the transformer encoder in Figure 6 (a) and the convolutional mask refinement network in Figure 6 (b). BLOCK is as defined in (10).

Fine branch (Section 3.2): The fine branch involves the composition of the TRANSFORMERENCODER features $\widetilde{F}$ with convolutional mask refinement network in (12), which produces $\widehat{F}$. Along the lines of (13), a binary classification head is subsequently applied as follows

$$S_{\text{fine}} \leftarrow \text{LCH}_1^{131}\left(\widehat{F}\right) \tag{16}$$

Here $S_{\text{fine}} \in [0,1]^{320\times320}$ is the high resolution *fine mask*. Therefore we denote the *fine branch* as

$$S_{\text{fine}} \leftarrow \text{FINEBRANCH}(X, \widetilde{F}). \tag{17}$$

where FINEBRANCH is given by

$$\text{CONVMASKREFINE} \rightarrow \text{LCH}_1^{131} \tag{18}$$

SEMPART (Section 3.4): The loss functions described in Section 3.2 and Section 3.3 are motivated by graph-based bi-partitioning of images based on deep semantic correspondences between regions as well as driven by graph total variation of the generated masks over the entire image. This results in high-quality self-supervised masks based on principles of normalized cut and guided super-resolution. We compute the corresponding loss functions in Section 3.4 to give us the eventual SEMPART loss in Algorithm 1.

The parameters of the transformer encoder, the convolutional mask refinement network, and the two binary classification heads are refined iteratively as per the loss $\mathcal{L}_{\text{SEMPART}}$. Note that this is an entirely unsupervised scheme where the DINO feature correspondences serve as the key source of self-supervision.

D. Supplementary material for Section 4

Architecture ablation comparison. Figure 7 demonstrates the architectural differences between SEMPART, and the ablations we compare with. In particular, as discussed in Section 4.4, we demonstrate the value of co-optimizing our *coarse* and *fine branches* (see Figure 7 (a)) as compared to only having the *fine branch* (see Figure 7 (c)) or having both branches trained independently (see Figure 7 (b)). Re-

[3]Note that this involves an average pooling step for aligning the spatial dimensions. See section on guided super-resolution in Section 3.2.

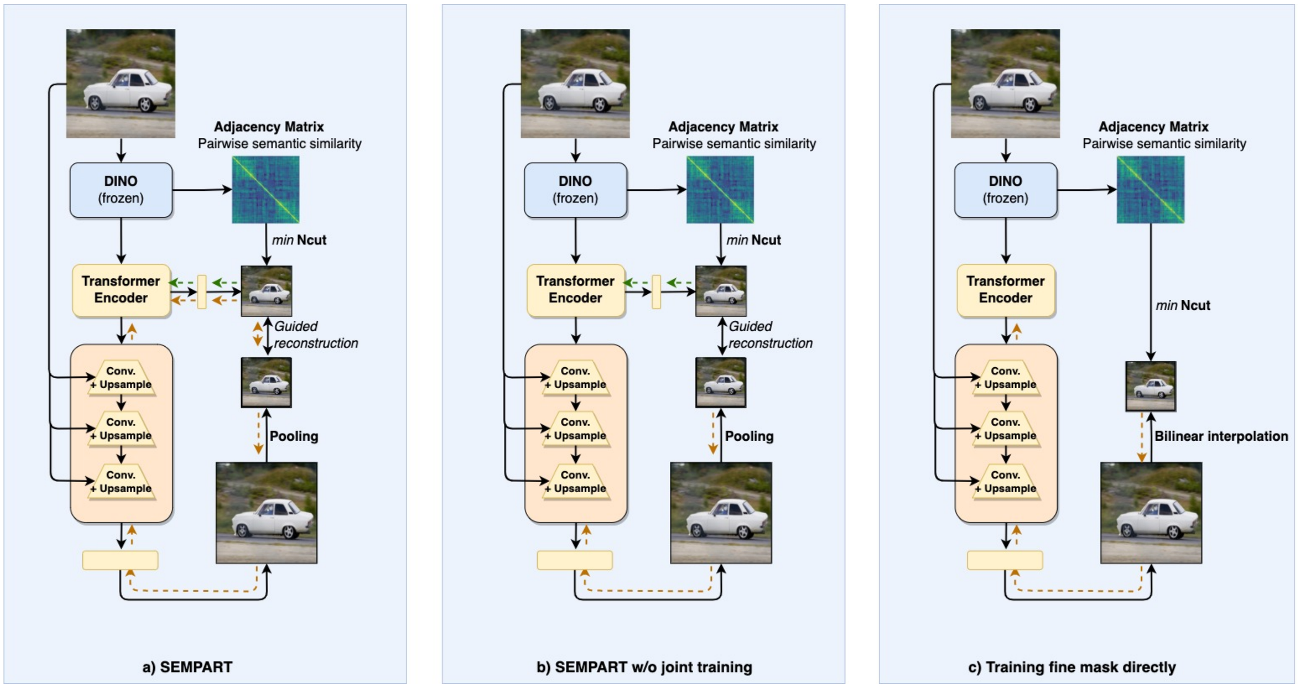

Figure 7 Comparison of SEMPART with ablations of its architecture in decreasing order of performance from (a) to (c) (see Table 4).

Algorithm 1 SEMPART

Input $X \in \mathbb{R}^{3\times 320\times 320}$ in RGB space
Output Loss $\mathcal{L}_{\text{SEMPART}}$

function $\text{LOSS}(X)$
  $F = \text{DINO}(X)$
  $\widetilde{F} = \text{TRANSFORMERENCODER}(F)$
  $S_{\text{coarse}} = \text{COARSEBRANCH}(\widetilde{F})$
  $S_{\text{fine}} = \text{FINEBRANCH}(X, \widetilde{F})$
  $\mathcal{L}_{\text{Ncut}} = \mathcal{L}_{\text{Ncut}}(F, S_{\text{coarse}})$ See (4)
  $\mathcal{L}_{\text{GTV-coarse}} = \mathcal{L}_{\text{GTV-coarse}}(F, S_{\text{coarse}})$ See Section 3.3
  $\mathcal{L}_{\text{SR}} = \mathcal{L}_{\text{SR}}(S_{\text{coarse}}, S_{\text{fine}})$[3] See (5)
  $\mathcal{L}_{\text{GTV-fine}} = \mathcal{L}_{\text{GTV-fine}}(X, S_{\text{fine}})$ See Section 3.3
  $\mathcal{L}_{\text{coarse}} = \mathcal{L}_{\text{Ncut}} + \lambda_{\text{GTV-coarse}}\mathcal{L}_{\text{GTV-coarse}}$
  $\mathcal{L}_{\text{fine}} = \lambda_{\text{GTV-fine}}\mathcal{L}_{\text{GTV-fine}}$
  $\mathcal{L}_{\text{joint}} = \lambda_{\text{SR}}\mathcal{L}_{\text{SR}}$
  $\mathcal{L}_{\text{SEMPART}} = \mathcal{L}_{\text{coarse}} + \mathcal{L}_{\text{fine}} + \mathcal{L}_{\text{joint}}$ See Section 3.4
  return $\mathcal{L}_{\text{SEMPART}}$
end function

Method	DUT-OMRON	DUTS-TE	ECSSD
w/o GTV coarse	$0.646\ (< 0.001)$	$\mathbf{0.749}\ (-)$	$0.848\ (< 0.001)$
w/o GTV fine	$0.637\ (< 0.001)$	$0.717\ (< 0.001)$	$0.818\ (< 0.001)$
train fine mask directly	$0.645\ (< 0.001)$	$0.738\ (< 0.001)$	$0.845\ (< 0.001)$
w/o joint training	$0.662\ (< 0.001)$	$0.743\ (< 0.001)$	$0.849\ (0.007)$
SEMPART-Fine	$\mathbf{0.668}$	$\mathbf{0.749}$	$\mathbf{0.855}$

Table 5 Ablations of SEMPART for saliency, using mIoU (p-value).

sults of the paired Wilcoxon signed-rank test [35] on the IoU metric, shown in Table 5, confirm the value of architectural choices, using significance level of 0.05.

As described in Section 4.4, Figure 7 (b) demonstrates that normalized cut loss only affects the transformer encoder and the *coarse branch*. In contrast, the gradients from the guided reconstruction only affect the *fine branch*. The gradients from the corresponding GTV losses also only affect the respective branches. In Figure 7 (c), however, the *coarse branch* is completely discarded, and the fine branch is utilized both for optimizing the expected normalized cut loss as well as the corresponding $\mathcal{L}_{\text{GTV-fine}}$ loss.

In our experiments (see Table 4), we observe that the performance in terms of the mean IoU of unsupervised saliency detection deteriorates consistently across all our evaluation datasets as we go from Figure 7 (a) to (b) to (c). This aligns with our intuition by demonstrating that not only is there value in separately inferring a *coarse mask* using the *coarse branch*, which effectively has the impact of a regularizer of the TRANSFORMERENCODER, but it is also beneficial to co-optimize the *fine* branch with the *coarse branch*.

Comparison with supervised methods. Table 6 compares the performance of SEMPART with recent state-of-the-art supervised methods [33, 58]. We show that using SEMPART masks for SelfMask training results in high quality masks outperforming the supervised U^2-NET on DUT-OMRON and DUTS-TE. However, a more recent supervised method

Method	OMRON*	D-TE*	ECSSD
SEMPART-Fine	0.668	0.749	0.855
SEMPART-Fine†	0.673	0.755	0.857
SELFMASK on SEMPART-Fine	0.698	0.749	0.850
U^2-NET[33]	0.693	0.733	0.878
SELFREFORMER[58]	0.744	0.830	0.900

†indicates that validation images were included during unsupervised training.

Table 6 We compare SEMPART variants with U^2-NET and SELFREFORMER both of which are supervised.

[58] still outperforms SEMPART by a significant margin.

We also observe that scaling the training set to also include the validation images improves the performance of SEMPART, indicated by SEMPART-Fine†.

Comparison with alternate backbones. Our experiments with alternate backbones in Table 7, indicates that the *degree of pixelation* (DoP), defined as the ratio of patch to image areas affects the performance. A larger ViT patch size is detrimental, and SSL features with lower DoP result in superior SEMPART saliency masks (Table 7 A, B vs. C, D). Nevertheless, the *fine mask* always outperforms its accompanying *coarse mask* by preserving high-frequency details.

	Backbone	Arch	Type	Input	DoP	OMRON	D-TE	ECSSD
A.	DINOv2 (2023)	ViT-S/14	Coarse	224^2	3.9e-3	0.460	0.539	0.659
B.			Fine	224^2	3.9e-3	0.523	0.598	0.717
C.			Coarse	560^2	6.25e-4	0.554	0.671	0.773
D.			Fine	560^2	6.25e-4	0.57	0.686	0.796
E.	DINO	ViT-S/16	Coarse	320^2	2.5e-3	0.573	0.640	0.766
F.			Fine	320^2	2.5e-3	0.596	0.656	0.793
G.		ViT-S/8	Coarse	320^2	6.25e-4	0.640	0.727	0.837
H.			Fine	320^2	6.25e-4	**0.668**	**0.749**	**0.855**

Table 7 SEMPART IoU (last three columns) for DINOv2 and DINO.

Hyperparameter sensitivity analysis. Figure 8 (A.1, A.2) show that the performance is typically robust to changes in λ_{SR} and $\lambda_{\text{GTV-coarse}}$ respectively. Figure 8 (A.3, B) show that the performance suffers with low and high $\lambda_{\text{GTV-fine}}$ values due to jaggedness and over-smoothing respectively.

Additional results. Figure 9, Figure 10 and Figure 11 present additional results for both SEMPART-coarse and -fine as well as also training SELFMASK+SEMPART-coarse and -fine as compared to TokenCut, MOVE, and the ground truth. The performance metrics in Table 1 indicate that the average performance of additionally training SELFMASK on SEMPART as pseudo masks results in an improvement of 3% and 3.5% in IoU and $\max F_\beta$ respectively for the DUT-OMRON dataset. At the same time, the gains are debatable for DUTS-TE and, in particular, for ECSSD, for which the performance deteriorates for the SELFMASK variant.

Across Figure 9, Figure 10, and Figure 11, the superiority of SEMPART over MOVE and TokenCut is a prevalent trend. As also seen previously in Figure 3, TokenCut, which is optimized on a per image basis, not only results in *coarse masks* that do not capture several high-frequency details but

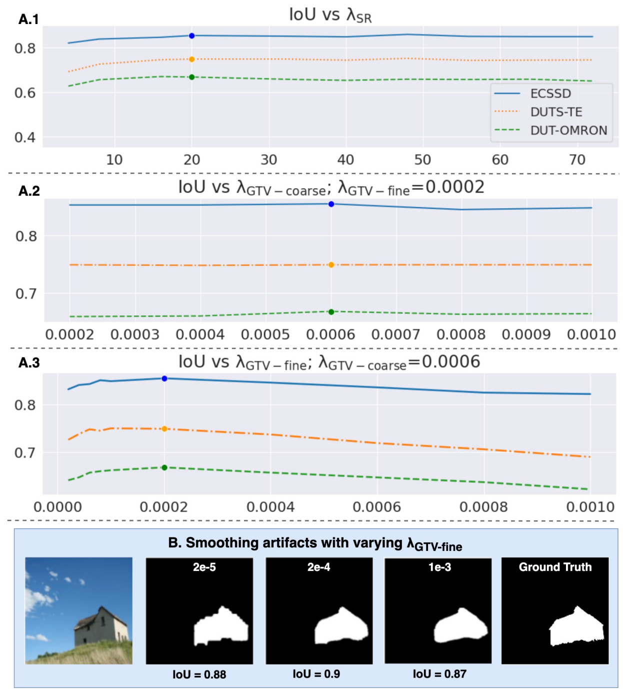

Figure 8 Hyperparameter sensitivity analysis of SEMPART-Fine.

can also select the incorrect object more often than its counterparts (see Figure 9 (I)) as well as under select the salient region (see Figure 9 (D, H), Figure 11 (C)).

On the other hand, MOVE outperforms TokenCut by generating more accurate and high-resolution masks based on the perception of *movability* of foreground objects. This heuristic outperforms previous state-of-the-art significantly, as demonstrated in [4]. However, we find that in addition to being noisy around the edges in most examples, it exhibits noisy artifacts both inside (see Figure 9 (G), Figure 10 (A, F), Figure 11 (B)) and outside (see Figure 9 (I), Figure 10 (E), Figure 11 (B, F))the visually salient regions. For the most part, MOVE can identify at least one of the salient objects. However, it seems likely that this heuristic also results in the over-selection of artifacts distinctly separated from the key salient object(s).

Compared to TokenCut and recent state-of-the-art MOVE, our method SEMPART and its SELFMASK variants signify a superior heuristic for unsupervised image bi-partitioning and a significantly better overlap with the ground truth saliency masks across all datasets. We also observe that the *fine mask* captures high-frequency details more accurately, especially at image boundaries than the corresponding jointly inferred *coarse mask*. The joint optimization involved in the SEMPART architecture is valuable towards image bi-partitioning without involving any post-inference processing. Therefore the inference times are a fraction of its counterparts and comparable with other methods that also learn a segmentation model, such as MOVE.

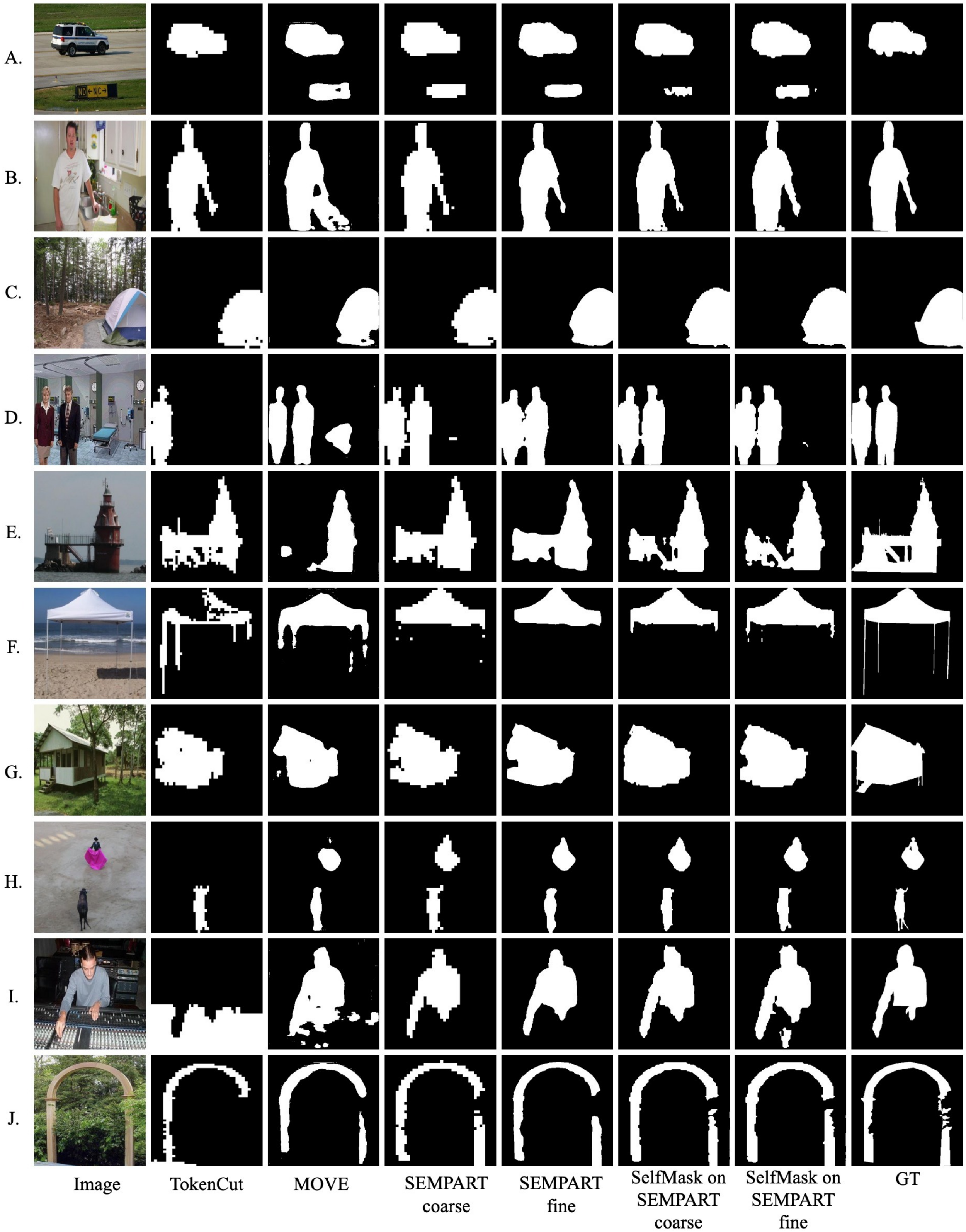

Figure 9 Additional examples on the DUT-OMRON [57] dataset.

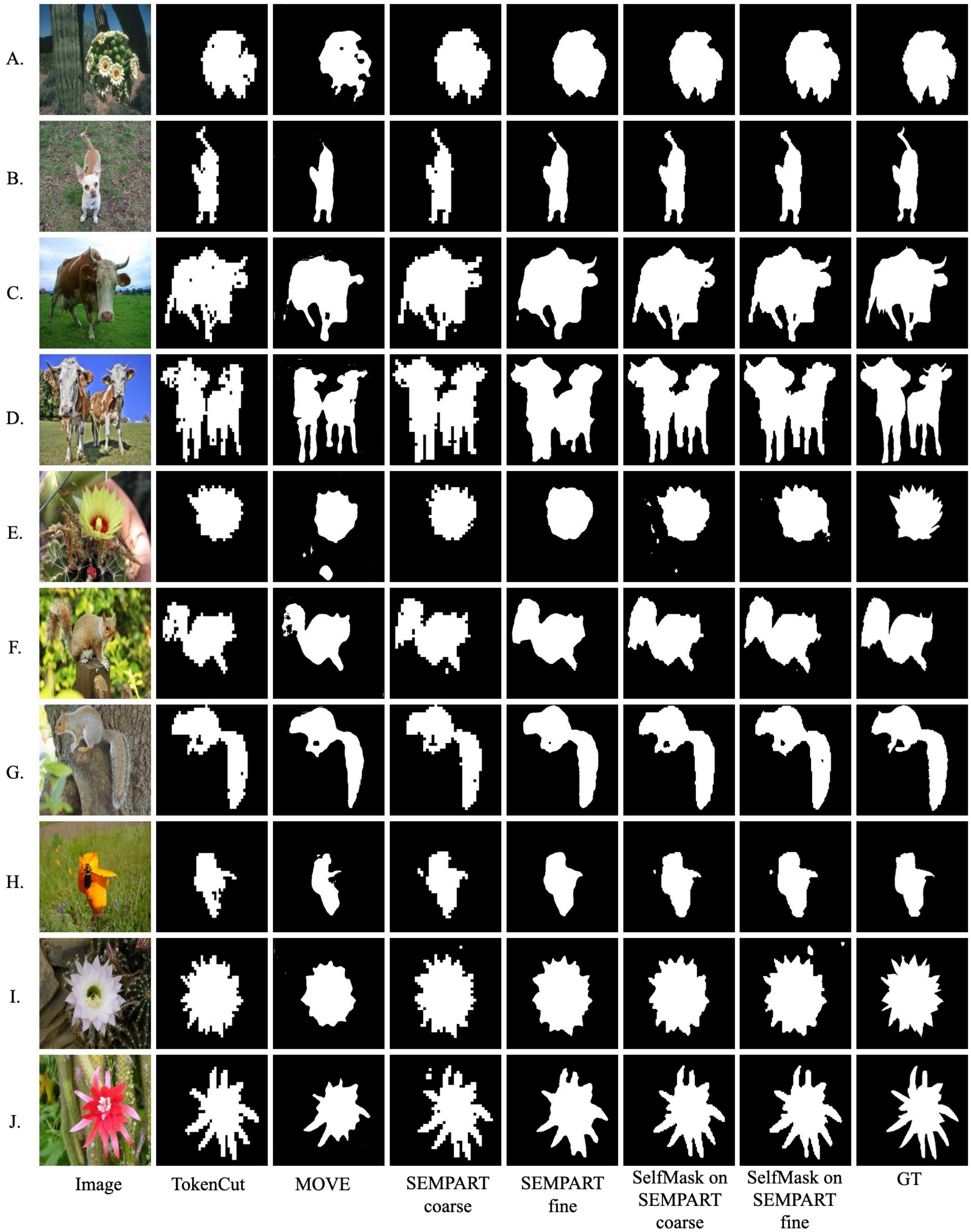

Figure 10 Additional examples on the ECSSD [38] dataset.

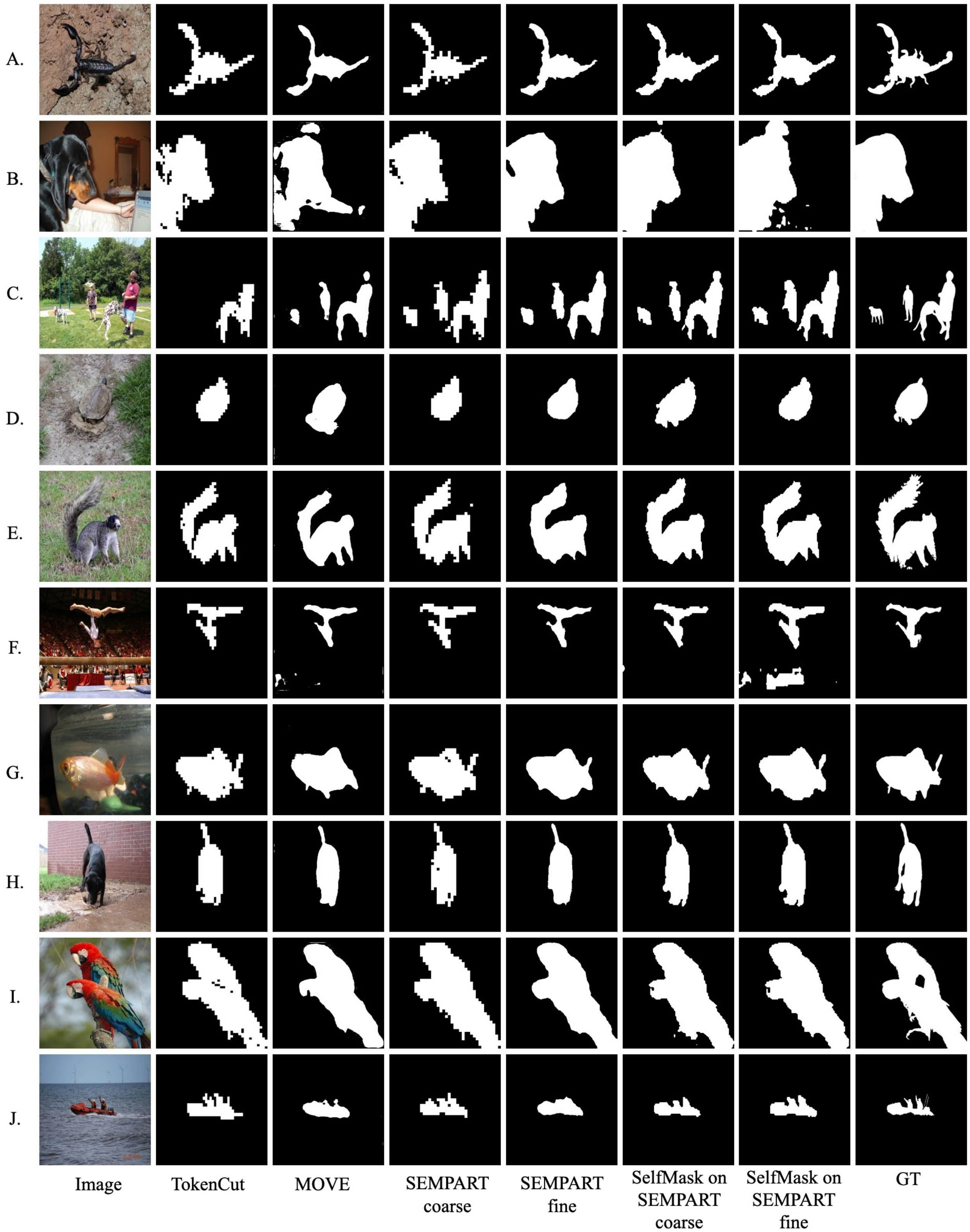

Figure 11 Additional examples on the DUTS-TE [49] dataset.

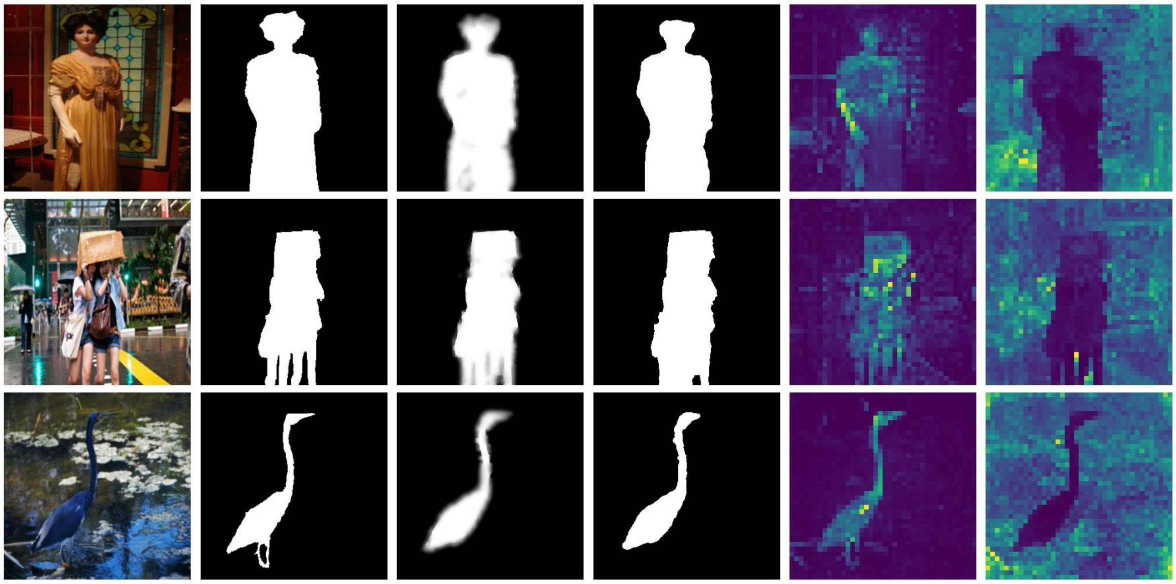

Figure 12 Attention map of the transformer encoder `[CLS]` token. The SEMPART attention map aligns with the background.

Attention map. The TRANSFORMERENCODER in (9) is further elaborated in Figure 6 (a). To get a better understanding of the reasoning process of SEMPART, we have looked at the average attention map across both heads for the `[CLS]` token of the TRANSFORMERENCODER in Figure 12. Interestingly we find that although the output of the TRANSFORMERENCODER for this particular token is discarded (see Figure 6 (a)), the corresponding attention map is insightful. This is because the `[CLS]` token is attended to by the remaining 40×40 patch tokens for generating $\widetilde{F}$ in (9). Therefore, the underlying `[CLS]` embeddings get leveraged for the $\widetilde{F}$ output. In particular, the attention map resonates with the background[4]. It reflects the clear distinction between an image's salient and non-salient regions. On the other hand, the DINO `[CLS]` token attention maps appear to attend to the foreground regions.

E. Ethical aspects

We benchmark our approach using publicly available datasets [49, 57, 38, 17, 18, 27]. Although our approach infers unsupervised partitions of images, SEMPART still inherits biases present in DINO [7], which was trained on ImageNet [13] without labels and in a self-supervised manner.

[4][42] adopted a heuristic that expands the mask from background seeds located first.

F. Future applications

The merits of SEMPART in generating high-quality masks at multiple resolutions can be particularly effective when applied to class-aware object detection, such as in [40]. More generally, SEMPART can also help improve search and recommendation systems [50] in applications where users seek to retrieve images of specific objects with the underlying assumption that the object under consideration will likely be prominent and in the foreground.

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, 10, 13

[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, 10, 17

[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. 17

[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, 10

[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, 17

[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, 17

[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, 17

[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, 12, 13

[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. 12

[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, 15, 17

[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, 17

[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, 17

[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, 16, 17

[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. 17

[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 colocalization 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, 14, 17

[58] Yi Ke Yun and Weisi Lin. Selfreformer: Self-refined network with transformer for salient object detection. *CoRR*, abs/2205.11283, 2022. 2, 12, 13

[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