BEYOND REPRESENTATION: SELF-SUPERVISED IMAGE ENCODERS TO SEGMENT ANYTHING

Anonymous authors
Paper under double-blind review

ABSTRACT

The Segment Anything Model (SAM) has been the state-of-the-art interactive image segmentation paradigm, which enables users to extract target objects with simple prompts. Despite its generalizability to unseen objects, it relies on large-scale annotated images for training. Meanwhile, Self-Supervised Learning (SSL) has yielded image encoders with strong visual representations. We therefore ask whether self-supervised image encoders alone can support "segment anything" without fully supervised training. We present SSL-SAM, a training-free framework based on on-the-shelf self-supervised vision encoders for interactive segmentation from points, boxes, masks, and scribbles. SSL-SAM converts each type of prompt into a support set of foreground and background respectively, and segments the target object by prototype-consistency scoring in the feature space of a frozen encoder. To deal with very sparse prompt (e.g., a single foreground click), we also propose an automatic support expansion strategy to achieve high performance while reducing the user's effort. Across 20 datasets with 5000+ classes, SSL-SAM outperforms the SAM family (SAM, SAM2, and SAM3) across multiple prompt modes, and achieves a 13.49% gain in mean Intersection over Union (mIoU) in box-only mode. These results show that self-supervised vision encoders can be adapted without retraining for promptable segmentation, which avoids the high cost of large-scale annotation and supervised training in the SAM models.

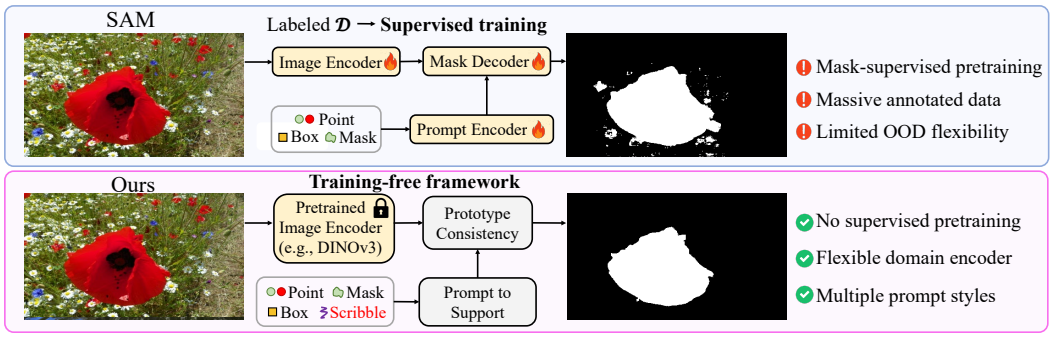


Figure 1: Illustration of difference between SAM and our SSL-SAM. SAM relies on a complex architecture and supervised training on large-scale manually annotated datasets. In contrast, SSL-SAM achieves promptable segmentation with only an image encoder pretrained by self-supervised learning, eliminating the need for large-scale manual annotation. Beyond SAM's point, box, and mask prompts, SSL-SAM additionally supports more flexible user interactions such as scribbles.

1 INTRODUCTION

Interactive image segmentation allows users to segment a target object with minimal manual efforts such as simply providing clicks, boxes, masks, or scribbles (Marinov et al., 2024). It is important not only for rapidly annotating image collections for training deep learning models, but also for im-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

proving segmentation robustness when automatic methods are unreliable, which is particularly acute in challenging settings, such as medical imaging and segmentation of previously unseen objects.

Though some classical interactive segmentation methods such as Graph Cuts (Boykov & Jolly, 2001), Random Walker (Grady, 2006), and GrabCut (Rother et al., 2004) have been proposed for decades, their reliance on low-level image features often leads to limited performance. Modern deep learning-based interactive methods trained with annotated images (Xu et al., 2016; Maninis et al., 2018; Sofiiuk et al., 2020; 2022) improved the performance, but have a limited support for different kinds of prompts and struggle at dealing with unseen objects (Chen et al., 2022; Liu et al., 2023; Luo et al., 2021). More recently, the Segment Anything Model (SAM) (Cheng et al., 2023) shifted the field towards general promptable segmentation by training an image encoder, a prompt encoder, and a mask decoder from large-scale annotated images with more than one billion masks, with the advantage of generalizability to previously unseen objects and multiple types of user inputs.

Despite the sucess of SAM and its variants (Xiong et al., 2024; Ma et al., 2024), constructing the required pre-training dataset with billions of masks incurs substantial human-annotation costs. In recent years, the development of self-supervised visual representation learning has achieved large progress, yielding image features with strong semantic and spatial structure (Caron et al., 2021; Siméoni et al., 2025). Recent studies have further shown that self-supervised pre-trained visual encoders are comparable to fully supervised ones when fine-tuned to different downstream tasks (Cuttano et al., 2026), and even support one-shot in-context segmentation under the guidance of a single support image (Li et al., 2024). These observations motivate us to ask an important question: Can large-scale supervised pre-training be avoided for promptable interactive segmentation, and is it possible to use on-the-shelf self-supervised visual encoders to achieve this goal?

We answer this question with SSL-SAM, a training-free framework that performs promptable segmentation by employing an on-the-shelf self-supervised image encoder. It converts user prompts into foreground and background support regions to obtain class prototypes, and then predicts the target object through prototype consistency scoring in feature space. As shown in Fig. 1, unlike SAM (Kirillov et al., 2023), SSL-SAM has a simpler architecture by removing the prompt encoder and the mask decoder, and therefore avoids supervised training. It also naturally allows multiple kinds of prompts, including scribble that is not well supported by SAM.

Our contributions are threefold. First, we introduce a general training-free framework for interactive image segmentation by directly employing on-the-shelf self-supervised pre-trained vision encoders, which naturally supports multiple kinds of prompts without supervised training and avoids large-scale annotation. Second, we propose a unified prototype-consistency inference mechanism based on support sets derived from different types of spatial prompts. Third, to deal with very sparse prompt, we propose prototype-guided support expansion to adaptively expand the support set for performance improvement while keeping minimal user interactions. We evaluated SSL-SAM on 20 datasets with more than 5000 classes across six prompt modes, and found that our method outperformed the SAM family by a large margin, especially under box-only, scribble and mask modes.

2 Related Work

2.1 Interactive Segmentation Methods

Traditional interactive segmentation methods mainly include scribble-based and bounding box-based paradigms, such as Graph Cuts (Boykov & Jolly, 2001), Random Walker (Grady, 2006), GrabCut (Rother et al., 2004), and Slic-Seg (Wang et al., 2016). Due to their image-specific training and low-level feature representations, these methods are sensitive to user inputs with limited performance in complex segmentation tasks. With the advantage of feature learning ability, deep neural networks have then been developed for interactive segmentation. Early methods combined clicks or distance transforms with Convolutional Neural Networks (CNNs) for this task (Xu et al., 2016; Maninis et al., 2018; Wang et al., 2019), and later methods improved iterative refinement, boundary quality, and efficiency through backpropagating refinement, transformers, and focal updates (Sofiiuk et al., 2020; 2022; Chen et al., 2022; Liu et al., 2023). Some methods have also explored to use bounding box and clicks for segmenting previously unseen objects with deep learning (Wang et al., 2018; Luo et al., 2021), but they are usually trained for specific datasets, domains, or prompt protocols, with limited generalizability to unseen targets, domains and prompt types.

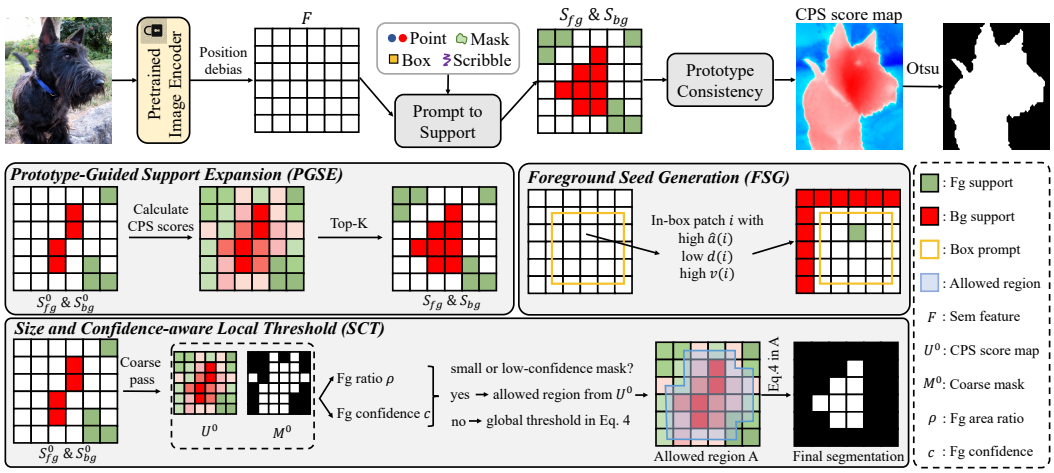


Figure 2: Overview of the proposed SSL-SAM framework. It converts different types of user interactions into a unified representation of foreground and background support sets, then segments the target object by prototype-consistency scoring between the support sets and the other regions with an on-the-shelf self-supervised vision encoder. The PGSE is used to expand the support set when user interactions are extremely sparse, and FSG generates the foreground seed automatically in box mode. SCT obtains an adaptive threshold for objects with small size of low confidence.

To deal with this problem, SAM (Kirillov et al., 2023) introduced a promptable and generalizable segmentation paradigm consisting of an image encoder, a prompt encoder, and a mask decoder, which was trained by supervised learning on 11 million images and more than one billion masks. Some subsequent works extended SAM for better performance on specific domains such as medical imaging (Cheng et al., 2023; Ke et al., 2023), higher efficiency, and supporting video segmentation (Xiong et al., 2024; Zhang et al., 2023; Ravi et al., 2025). However, their generalizability is obtained at the high expense of large-scale annotation and supervised training, which limits transferring the paradigm to boarder domains. In addition, the supervised training with some specific kinds of prompts limit the users to give more flexible interactions that may be unseen during training. For example, SAM has a limited support for scribbles that are widely used in traditional methods.

2.2 Self-Supervised Visual Encoders for Segmentation

Self-Supervised Learning (SSL) avoids human annotation cost by learning transferable representations from unannotated images. The widely used SSL strategies include contrastive learning (Chen et al., 2021), Masked Auto Encoder (MAE) (He et al., 2022; Bao et al., 2021), and self-distillation (Caron et al., 2021; Oquab et al., 2024). Recent studies have shown that SSL-based vision encoders can be fine-tuned to achieve high performance on different downstream tasks including image classification (Siméoni et al., 2025), segmentation (Yang et al., 2025), and salitent object detection (Wang et al., 2022). What's more, Cuttano et al. (2026) show that SSL vision encoders can support in-context segmentation (Cuttano et al., 2026). Differently from these works, we explore the feasibility of using on-the-shelf SSL vision encoders for training-free and generalizable interactive segmentation, which has been rarely investigated in the literature.

3 Method

Given an image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$, a user prompt $\mathbf{P}$, and a frozen pretrained image encoder E (a vision Transformer), SSL-SAM predicts a binary target mask $\mathbf{M} \in \{0,1\}^{H\times W}$. The prompt $\mathbf{P}$ may contain points, a bounding box, a coarse mask, scribbles, or their combination. For the input image $\mathbf{I}$, the feature map obtained by E is denoted as:

$$\mathbf{F} = E(\mathbf{I}) \in \mathbb{R}^{H'\times W'\times C}, \quad H' = H/p, \quad W' = W/p, \tag{1}$$

where p is the down-sampling factor in the encoder, and C is the patch-level feature dimension after removing the spatial embedding in vision transformers (Cuttano et al., 2026). We then derive a foreground support set and a background support set from $\mathbf{P}$ (Sec. 3.2) to calculate class-wise prototypes that guide the subsequent segmentation task (Sec. 3.1), which is detailed in the following.

3.1 Prototype-Based Segmentation

Let S_{fg} and S_{bg} denote the foreground and background support set of pixels derived from a certain prompt for the testing image, respectively. The foreground and background prototypes are:

$$\mathbf{f}_{fg} = \frac{1}{|S_{fg}|} \sum_{\mathbf{x} \in |S_{fg}|} \mathbf{F}(\mathbf{x}), \qquad \mathbf{f}_{bg} = \frac{1}{|S_{bg}|} \sum_{\mathbf{x} \in S_{bg}} \mathbf{F}(\mathbf{x}), \tag{2}$$

where $\mathbf{x}$ denotes a pixel in the feature map. We then calculate the pixel's similarity to $\mathbf{f}_{fg}$ and $\mathbf{f}_{bg}$ respectively, and obtain a Contrastive Prototype-Similarity (CPS) score as:

$$U(\mathbf{x}) = (\mathbf{F}(\mathbf{x}))^\top \mathbf{f}_{fg} - (\mathbf{F}(\mathbf{x}))^\top \mathbf{f}_{bg}, \tag{3}$$

where the similarity between $\mathbf{F}(\mathbf{x})$ and a prototype is defined by their dot product. Note that in some cases (e.g., only foreground clicks are provided), S_{bg} may be empty and $\mathbf{f}_{bg}$ is not available, and we simply define the CPS score as: $U(\mathbf{x}) = (\mathbf{F}(\mathbf{x}))^\top \mathbf{f}_{fg}$. Let τ denote a threshold of the CPS score, the segmentation output at the feature map's resolution is denoted as $M' = \mathbb{I}[U > \tau]$. Considering that a fixed threshold is not adaptive to different testing images, we use Otsu optimization to automatically find the optimal threshold τ^* by minimizing intra-class variances in the CPS map:

$$\tau^* = \arg\min_{\tau} \left[\omega_{fg}(\tau)\sigma^2_{fg}(\tau) + \omega_{bg}(\tau)\sigma^2_{bg}(\tau)\right], \tag{4}$$

where $\omega_{fg} = |M'|/(H' \times W')$ and $\omega_{bg} = 1 - \omega_{fg}$ are the ratio of foreground and background region in M', respectively, and $\sigma^2_{fg}(\tau)$ and $\sigma^2_{bg}(\tau)$ are CPS variance in the two regions under threshold τ, respectively. The mask with the optimal threshold τ^* is then upsampled to the original image resolution to obtain the final segmentation mask M. As the global Otsu threshold may lead to over-segmentation when the target is very small, we propose Size and Confidence-aware Local Threshold (SCT) to deal with this issue in Sec. 3.3.

3.2 Prompt to Support Transform

To leverage the general prototype-based interactive segmentation framework in Sec. 3.1, we transform a prompt into a set of foreground support S_{fg} and a set of background support S_{bg}. For click and scribble inputs, S_{fg} and S_{bg} are simply defined as the foreground and background seeds, respectively. For mask prompt, we use the inner region of the eroded mask and the outside of the dilated mask as S_{fg} and S_{bg}, respectively. For the click prompt mode, to deal with very sparse clicks (e.g, a single click for the foreground), we propose Prototype-Guided Support Expansion (PGSE) to expands the support set. For bounding box prompt, the outside of the box is used as S_{bg}, and S_{fg} is obtained by our automatic Foreground Seed Generation (FSG) strategy.

Prototype-Guided Support Expansion (PGSE). Under very sparse prompts with extremely small support sets, especial in the scenario of single click for each class, the prototypes obtained by Eq. 2 may not be representative enough and limit the segmentation performance. To deal with this issue, we propose PGSE that automatically expands the support sets based on context of the target image. Let S^0_{fg} and S^0_{bg} denote the raw support sets, and U^0 be the corresponding CPS score map, we add K pixels with the hightest and lowest CPS scores to S^0_{fg} and S^0_{bg}, respectively. Let Ω denotes the entire image space, the support expansion results are:

$$\mathcal{S}_{fg} = \mathcal{S}^0_{fg} \cup \mathrm{TopK}_K\left(U^0, \Omega \setminus \mathcal{S}^0_{bg}\right), \quad \mathcal{S}_{bg} = \mathcal{S}^0_{bg} \cup \mathrm{BottomK}_k\left(U^0, \Omega \setminus \mathcal{S}^0_{fg}\right), \tag{5}$$

where TopK_K and $\mathrm{BottomK}_K$ select the K points with the highest and lowest CPS scores in $\Omega \setminus \mathcal{S}^0_{bg}$ and $\Omega \setminus \mathcal{S}^0_{fg}$, respectively. Then, the expanded support sets $\mathcal{S}_{fg}$ and $\mathcal{S}_{bg}$ are used to obtain the final segmentation output following Sec. 3.1.

Foreground Seed Generation (FSG) for box mode. In the bounding box mode, the user does not explicitly assign any pixels with the foreground label. To deal with this, we propose FSG to

automatically obtain a reliable foreground seed by combining three complementary cues: attention score, dissimilarity to background prototype, and spatial centrality. Firstly, attention score provides a box-conditioned measure of object relevance, and highlights semantically coherent patches that contribute strongly to the box-level representation. Let J denote the number of spatial tokens in the bounding box, G be the number of heads in the transformer layer. The key in the g-th head of the j-th token is denoted as $\mathbf{k}_j^{(g)}$. The attention score of the i-th token in the bounding box is defined as the class-token-to-patch attention in the last Transformer layer of the image encoder E:

$$a(i) = \frac{1}{G}\sum_{g=1}^{G}\left[\mathrm{softmax}_j\left(\frac{\mathbf{q}_{\mathrm{CLS}}^{(g)}\cdot\mathbf{k}_j^{(g)}}{\sqrt{d_g}}\right)\right]_i, \tag{6}$$

where d_g is the head dimension, and $\mathbf{q}_{\mathrm{CLS}}^{(g)}$ denote the class-token query in the g-th head. The attention scores are then normalized to [0,1] based on the minimal and maximal values in the bounding box. Secondly, dissimilarity to background prototype is denoted as $d(i) = 1 - sim(\mathbf{f}_i, \mathbf{f}_{bg})$, where $sim(\cdot,\cdot)$ is the cosine similarity. A lower $d(i)$ value indicates a higher probability for being the foreground. Thirdly, the centrality score is defined as $v(i) = 1 - \mathrm{dis}(i)/\mathrm{dis}_{max}$, where $\mathrm{dis}(i)$ is the Euclidean distance to bounding box with dis_{max} being the maximal distance value. FSG selects the patch with the highest combined score as the foreground seed in the bounding box:

$$i^* = \arg\max_{i\in\mathcal{R}_{\mathrm{box}}}\left[\hat{a}(i) + d(i) + v(i)\right]. \tag{7}$$

where $\hat{a}(i)$ is the normalized attention score, and $\mathcal{R}_{\mathrm{box}}$ denotes the bounding box region. By comprehensively considering the three scores from different aspects, the foreground seed selection has a higher robustness than using each score independently. As a single foreground seed selected from the bounding box is still sparse, we apply the PGSE strategy to expand the foreground support set in the bounding box for segmentation.

3.3 Size and Confidence-aware Local Threshold (SCT)

Although Eq. 4 is generally effective in most cases for segmentation, in some challenging cases, some background region may have moderate CPS scores due to ambiguity between foreground and background, especially in scenarios with low contrast and small object size, where the Otsu optimization in the entire image space can easily lead to over-segmentation. To address this issue, we propose SCT for small objects or uncertain segmentation, which applies Eq. 4 to calculate the optimal threshold in a local allowed region $\mathcal{A}$. First, with the raw supports $\mathcal{S}_{fg}^0$ and $\mathcal{S}_{bg}^0$, the corresponding CPS score map U^0 and a coarse mask M^0 are obtained following Sec. 3.1. We estimate the relative size of the object by the foreground area ratio ρ and estimate the segmentation confidence c by the average CPS score in the foreground region:

$$\rho = \frac{\sum_{\mathbf{x}\in\Omega} M^0(\mathbf{x})}{|\Omega|}, \qquad c = \frac{\sum_{\mathbf{x}\in\Omega} M^0(\mathbf{x})U^0(\mathbf{x})}{\sum_{\mathbf{x}\in\Omega} M^0(\mathbf{x})}, \tag{8}$$

where Ω is the entire image space. When ρ and c are above their corresponding thresholds respectively, $\mathcal{A}$ is defined as the entire image space. Otherwise, it is defined as the high-CPS area:

$$\mathcal{A} = \left\{\mathbf{x}\in\Omega \mid U^0(\mathbf{x}) \geq u_q\right\}. \tag{9}$$

where u_q is a threshold determined by the q-th percentile in the CPS map.

Based on Sec. 3.2, when S_{fg}^0 is expanded for cases where it is extremely small (i.e., a single foreground click, or one generated foreground seed for box mode), after getting the new CPS score map U based on expanded support sets $\mathcal{S}_{fg}$ and $\mathcal{S}_{bg}$, we apply Eq. 4 in the allowed region $\mathcal{A}$ to obtain the final segmentation mask with a locally adaptive threshold, and set outside of $\mathcal{A}$ as the background.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluated our method on 20 datasets with more than 5000 classes from six different domains: 1) three general-purpose object segmentation datasets including COCO-MVal, SBD,

Table 1: Summary of the 20 datasets. N/A indicates no official category number.

Task domains	Dataset	images	Categories	Fg ratio (%)	Representative categories
General object seg	COCO-MVal (Lin et al., 2014)	5000	80	17.47	Person, bicycle, chair
	SBD (Hariharan et al., 2011)	11355	20	20.34	Aeroplane, bottle, bus
	VOC2012 (Everingham et al., 2010)	2913	20	17.95	Bicycle, bird, boat, cow
Salient object seg	DUT-OMRON (Yang et al., 2013)	5168	749	15.50	People, animals
	DUTS-TE (Wang et al., 2017)	5019	1152	16.01	Animals, people, vehicles
	DUTS-TR (Wang et al., 2017)	10553	1152	29.16	Tiger, wild boar, mushroom
	ECSSD (Shi et al., 2015)	1000	310	22.89	Animals, people
	MSRA10K (Liu et al., 2010)	10000	100+	21.74	Single people, products
	PASCAL-S (Li et al., 2014)	850	N/A	33.59	people, vehicles, animals
	SOD (Movahedi & Elder, 2010)	300	N/A	28.97	Elk, starfish, horse
	S3OD (Kupyn et al., 2026)	139981	1676	13.72	Synthetic reflex cameras
Specialized seg	CUB-200 (Wah et al., 2025)	11788	200	16.66	warblers, woodpeckers
	OxfordPet (Parkhi et al., 2012)	7390	37	45.18	Bengal, Persian, Beagle
	F-Beasts (Ma et al., 2023)	250	20	28.70	Billywig, Graphorn
Challenging seg	CAMO (Le et al., 2019)	1250	500+	12.93	Camouflaged animals, insects
	DIS5K (Qin et al., 2022)	5000	758	19.14	Aircraft, furniture
Video object seg	DAVIS16 (Perazzi et al., 2016)	3455	50	7.98	Moving animals, people
Medical image seg	ISIC (Codella et al., 2019)	2,594	1	24.77	Melanoma, nevus
	Kvasir (Jha et al., 2019)	1000	1	16.83	Gastrointestinal polyps
	TN3K (Gong et al., 2023)	3493	1	18.13	Thyroid nodules

and VOC2012; 2) eight salient object segmentation datasets including DUT-OMRON, DUTS-TE, DUTS-TR, ECSSD, MSRA10K, PASCAL-S, SOD, and S3OD; 3) three specialized object segmentation datasets (CUB-200, OxfordPet, and F-Beasts); 4) two datasets for challenging object segmentation (CAMO and DIS5K); 5) one video object segmentation dataset (DAVIS16); and 6) three medical image segmentation datasets (ISIC, Kvasir, and TN3K) in different modalities. Table 1 summarizes the image number, average foreground area ratio and representative target categories of each dataset. For each dataset, we randomly sampled 200 images for evaluation. We also sampled another 200 images from OxfordPet and DAVIS16 respectively for hyperparameter study.

Implementation and Evaluation Details. Our method used DINOv3 ViT-L/16 as the frozen encoder in the main experiments. Six prompt modes were used for evaluation: 1) 1fg that means a single foreground click, 2) 3fg + 3bg where three foreground and three background clicks are provided, 3) box mode, 4) box + 1fg + bg, 5) mask mode, and 6) scribble mode. The prompts were automatically generated by a script based on the ground truth mask with some perturbations. PGSE with FSG and SCT were only applied to 1fg, 3fg + 3bg and box modes, and not to the other prompt modes due to the relatively large size of raw support sets. For PGSE, we used $K = 10$ for point and box modes, respectively. For SCT, u_q was set as the 90-th percentile of U^0 in Eq. 9.

We compared SSL-SAM with SAM (Kirillov et al., 2023), SAM2 (Ravi et al., 2025), and SAM3 (Carion et al., 2025) under the same set of prompts. For box mode, we also included two other competitive methods GrabCut (Rother et al., 2004) and TokenCut (Wang et al., 2022) that are designed for box prompts. As SAM does not support scribble prompt natively, we view the scribble as point prompt for comparison. We also included two scribble-based interactive segmentation methods Random Walker (Grady, 2006) and GraphCut (Boykov & Jolly, 2001). For quantitatively evaluation, we reported mean Intersection over Union (mIoU) averaged across different target classes. No post processing was implemented for all the methods for a fair comparison.

4.2 Comparison with state-of-the-art

Fig. 3 shows quantitative evaluation results in terms of mIoU on the 20 datasets and six prompt modes. Our methods generally outperforms the existing methods in all the prompt modes. In the 1fg mode, SSL-SAM achieved an mIOU of 66.48%, outperforming SAM (47.37%), SAM2 (61.29%) and SAM3 (64.15%). Among the 20 datasets, SSL-SAM was superior to SAM, SAM2 and SAM3 on 20, 16 and 12 datasets, respectively. In the 3fg+3bg mode, the mIoU of SSL-SAM increased to 69.80%, and exceeded SAM, SAM2 and SAM3 by 0.26, 3.37 and 6.98 points, respectively. Under box-only prompting, SSL-SAM obtained an average mIOU of 72.93%, outperforming SAM, SAM2, and SAM3 by 14.06, 14.83, and 11.81 percentage points, respectively. In the box + 1fg + 1bg mode,

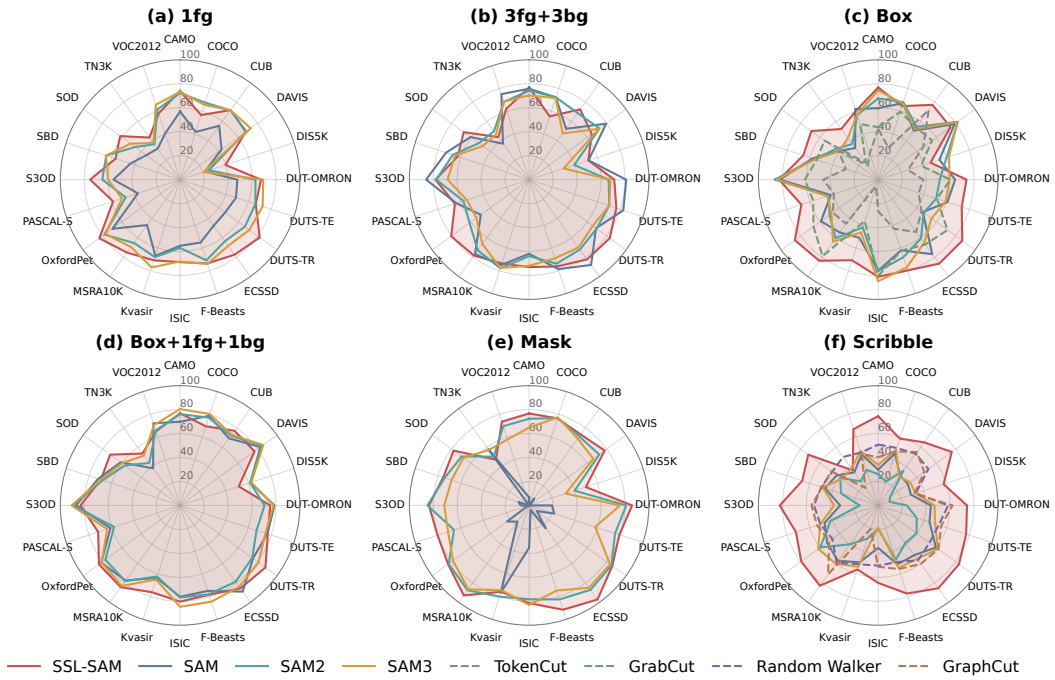


Figure 3: Quantitative comparison across six prompt modes and 20 datasets. Dataset-level mIoU is reported. The dashed lines show additional box or scribble-based existing methods for comparison.

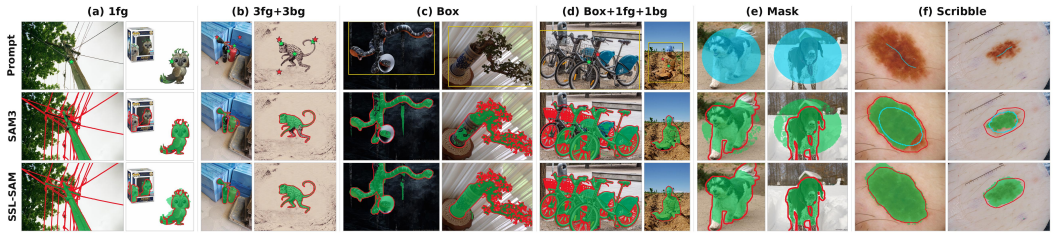


Figure 4: Representative comparisons between SSL-SAM and the strongest baseline SAM3. Red contours indicate ground-truth masks, cyan contours in (f) indicate the random walker prediction, and green regions indicate predicted masks. More results are shown in the Supplementary material.

SSL-SAM achieved an mIoU of 74.82%, and outperformed SAM and SAM2 by 2.45 and 3.96 percentage points, respectively. Despite the slightly lower mIoU (a gap of 0.41 percentage points) compared with SAM3, our method outperformed SAM3 on 11 datasets. Note that the mask and scribble modes are not well supported by SAM. Our method obtained an average mIoU of 78.30% in the mask mode, which outperformed SAM, SAM2, and SAM3 by 61.76, 4.95, and 13.72 percentage points, respectively. In the scribble mode, the SAM family obtained a very low performance (40.19% at most), while our method kept a high performance with average mIoU of 70.44%, which also largely outperformed GraphCut and Random Walker. These results showed better generalizability of SSL-SAM on different datasets and prompt modes than the existing methods.

Fig. 4 presents qualitative comparison between SAM3 and our method with different prompt modes and datasets. In the scribble mode, we also included Random Walker for comparison (cyan contour). In the 1fg mode, SAM3 tends to obtain under-segmentation for targets with narrow parts, irregular boundaries and multiple instances, while our method can better deal with these cases, and the same observation is found in the 3fg + 3bg mode. In the box and box + 1fg + 1bg mode, both methods achieve good results, but SAM3 has more obvious under- and over-segmentation when the target has

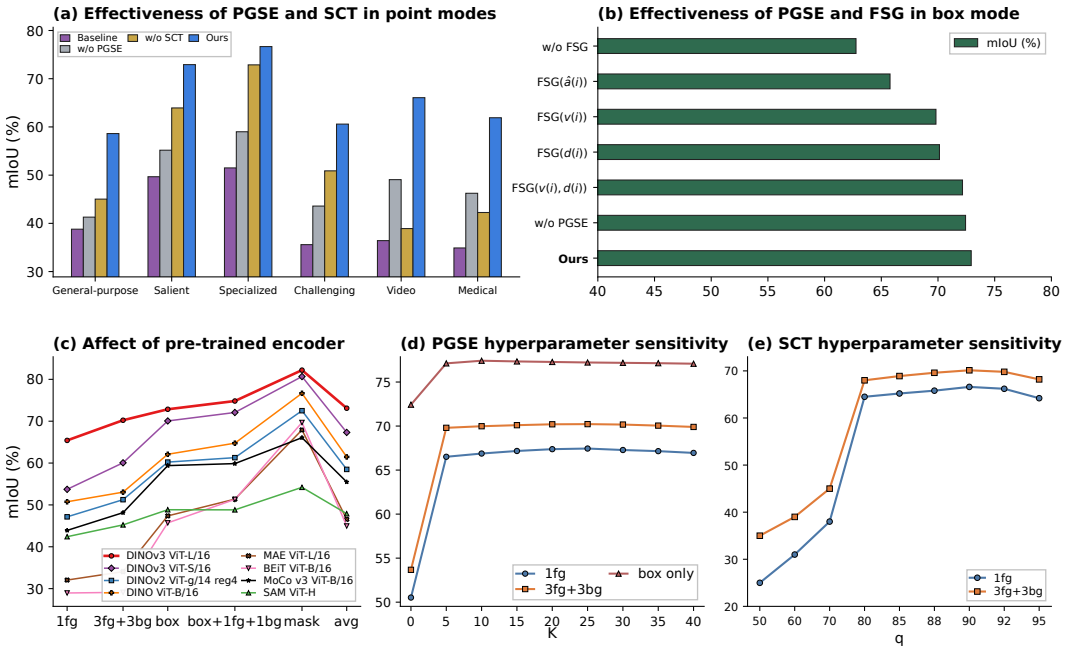


Figure 5: Ablation study results. (a) Contribution of the SCT and PGSE, averaged over the 1fg and 3fg+3bg prompt settings. (b) Effect of PGSE and FSG in the box mode. (c) Results of using different pre-trained encoders in our framework. (d) Effect of K in PGSE for 1fg, 3fg+3bg, and box-only prompts. (e) Sensitivity of the q-th percentile in SCT for 1fg, 3fg+3bg prompts.

a heterogenous appearance and complex boundary. In the mask mode, when the prompt mask has a relatively poor quality, our method obtains a higher robustness than SAM3. In addition, when only a foreground scribble is provided, Random Walker and SAM3 have an obvious under-segmentation, while our method better extracts the entire lesion in Fig.3(f). These examples show that SSL-SAM can flexibly deal with different kinds of prompts, and is able to convert sparse or coarse prompts into coherent masks while preserving complex shapes and detailed boundaries.

4.3 Ablation Study

Effectiveness of PGSE and SCT in point modes. As our PGSE and SCT are proposed to deal with sparse prompts (one or few points) or small objects, we first evaluated their effectiveness in 1fg and 3fg + 3bg modes. The baseline was the naive method in Sec. 3.1, and then we added PGSE and SCT gradually. Fig. 5(a) shows a comparison of these variants on the six domains of datasets, where the average performance of the two point modes is reported. It can be observed that the baseline obtained the worst performance, and the complete strategy achieved the highest average mIoU (69.15%) across all the dataset. Removing the PGSE and SCT decreased the mIoU to 46.22% and 41.95%, respectively. Notably, removing SCT leads to a substantial performance degradation on the video group. This is mainly because the video datasets in our experiments contain a large number of small targets, for which SCT effectively constrains the mask prediction and helps preserve the target regions. These results show that the PGSE and SCT are effective to obtain robust performance under point prompts that are very sparse in the image.

Effectiveness of PGSE and FSG in box mode. We further conducted ablation study of PGSE and FSG in the box mode. As shown in Fig. 5(b), our complete strategy achieved an average mIoU of 72.93% across the testing datasets, and removing PGSE (i.e., the foreground support set only has a single seed generated by FSG) reduced it by 2.98 percentage points. Removing FSG (keeping foreground support empty and using background dissimilarity for segmentation) caused a 10.56% decrease. In addition, we evaluated four variants of FSG that use parts of the three cues for generating the foreground seed: attention score $\hat{a}(i)$ only, dissimilarity to background $d(i)$ only, centrality

$v(i)$ only, and $v(i) + d(i)$. These variants obtained a reduced mIoU of 65.77%, 70.13%, 69.82%, and 72.16%, respectively, and they all underperformed the full version of FSG.

Affect of pre-trained encoder. We analyzed the performance of different pre-trained encoders in our SSL-SAM framework in Fig. 5(c). DINOv3 ViT-L/16 achieved the best performance in all the prompt modes, due to the high token-level discrimination ability from Gram anchoring-based self-supervised pre-training (Siméoni et al., 2025). DINO and DINOv2 also outperformed the other SSL-based pre-trained encoders including MAE ViT-L/16 and MoCo v3 ViT-B/16. In addition, these SSL-based encoders outperformed BEiT ViT-B/16 pre-trained by supervised learning. What's more, the SAM's image encoder learned by large-scale supervised training performed worse than the DINO-based encoders when used in our framework. These results suggest that our SSL-SAM is compatible to different kinds of SSL-based vision encoders, and they are are better than supervised image encoders in our framework, mainly due to that the supervised one has biased features (Zhang et al., 2025) that reduce generalizability and SAM encoder has a low semantic discrimination ability.

Hyper-parameter sensitivity. Fig. 5(d) shows the effect of hyper-parameter K for support expansion. Under three sparse prompt settings (1fg, 3fg + 3bg, and box with FSG), the expanded support sets with $K \geq 5$ largely outperform $K = 0$ (i.e., no expansion). In addition, the results were not sensitive to K values when K ranged from 5 to 40, so we set $K = 10$ for both point and box modes. Fig. 5(e) shows the effect of the q-th percentile for defining the allowed region in SCT under the 1fg and 3fg+3bg prompt modes. The performance gradually improved as q increased from 50 to 70, followed by a sharp increase at $q = 80$ and then reaching a plateau. This result indicates that when $q \geq 80$, SCT can effectively alleviate the limitations of global Otsu thresholding in images with small targets and low foreground-background contrast. The performance is stable when q ranges from 80 to 95, so we set $q = 90$ in the configuration. The effect of two other hyper-parameters (thresholds for ρ and c in SCT) is presented in the Supplementary material, which shows that our method is not sensitive to these hyper-parameters as well.

5 Discussion and Conclusion

The success of SAM tells the community that large-scale annotation and supervised training are crucial for generalizable interactive segmentation. Our SSL-SAM challenges this existing practice by demonstrating that on-the-shelf self-supervised pre-trained vision encoders already contains enough semantic representations to support promptable segmentation, which leads to zero human annotation cost and avoids supervised learning. On different types of datasets, SSL-SAM can achieve even better results than the SAM families, and it naturally supports different kinds of prompts with a unified representation of foreground and background support sets. As shown in Fig. 5(e) and Fig. 3(f), our training-free framework is compatible with different kinds of vision encoders and more flexible user inputs that may not be well supported by SAM. Our framework is challenged by very sparse prompts where the support set may lead to unstable class prototypes, and we deal with this by the PGSE strategy and automatically generating the reliable foreground seed in box mode. In addition, the SCT strategy plays an important role in dealing with small objects where applying the Otsu thresholding in the entire image space can easily lead to over-segmentation, as shown in Fig. 5(a). These strategies increase adaptability of our prototype-based framework, and are important for achieving robust results in various testing image domains.

One limitation of our method is that it faces the same patch-level prediction problem as SAM, i.e., the mask has a lower resolution than the input image. A post-precessing such as Conditional Random Field (CRF) can help to alleviate this problem and obtain higher boundary accuracy. In addition, our method has not considered the 'everything' mode for automatic segmentation yet. In the future, it is of interest to combine context-aware automatic prompt generation with our unified support sets and prototype-based segmentation framework with SSL vision encoders.

In conclusion, we introduced SSL-SAM, a training-free framework that turns an on-the-shelf self-supervised image encoder into a promptable segmentation tool through prompt-to-support transform and prototype-consistency scoring. Across 20 datasets and multiple prompt modes, SSL-SAM matches or even outperforms the supervised counterparts in the SAM-family, with the strongest gains in box and scribble modes, leading to more generalizable and flexible interactive segmentation, which sets a new paradigm for high-quality image annotation with minimal user interactions.

AI USE STATEMENT

In this work, we used generative AI tools solely to assist with language editing, including improving grammar, clarity, readability, and overall academic phrasing of the manuscript. We did not use generative AI tools to generate or substantively develop the research ideas, methodology, experimental design, data analysis, results, or conclusions presented in this work. All AI-assisted text was reviewed and revised by the authors to ensure its accuracy, consistency with the research, and compliance with the intended meaning. The authors take full responsibility for the final content of this work, including all text, claims, and artifacts.

REPRODUCIBILITY STATEMENT

We have taken several measures to facilitate the reproducibility of our work. The supplementary material provides extensive qualitative visualizations and additional experimental results across different datasets and prompt types, offering further evidence of the proposed framework's behavior under diverse segmentation settings. Detailed descriptions of the datasets, experimental protocols, evaluation metrics, and implementation details are provided in the main paper and supplementary material. Upon acceptance, we will publicly release the complete source code and relevant configuration files to enable independent reproduction of our experiments.

REFERENCES

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

Yuri Y Boykov and M-P Jolly. Interactive graph cuts for optimal boundary & region segmentation of objects in nd images. In *ICCV*, volume 1, pp. 105–112, 2001.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025.

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

Xi Chen, Zhiyan Zhao, Yilei Zhang, Manni Duan, Donglian Qi, and Hengshuang Zhao. Focalclick: Towards practical interactive image segmentation. In *CVPR*, pp. 1300–1309, 2022.

Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *ICCV*, pp. 9620–9629, 2021.

Junlong Cheng, Jiayuan Ye, Zhongqiang Deng, Jianpin Chen, Tianbin Li, Haoyu Wang, Yanzhou Su, Ziyan Huang, Jilong Chen, Lei Jiang, et al. Sam-med2d. *arXiv preprint arXiv:2308.16184*, 2023.

Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*, 2019.

Claudia Cuttano, Gabriele Trivigno, Christoph Reich, Daniel Cremers, Carlo Masone, and Stefan Roth. Insid3: Training-free in-context segmentation with dinov3. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21638–21648, 2026.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2): 303–338, 2010.

Haifan Gong, Jiaxin Chen, Guanqi Chen, Haofeng Li, Guanbin Li, and Fei Chen. Thyroid region prior guided attention for ultrasound segmentation of thyroid nodules. *Computers in biology and medicine*, 155:106389, 2023.

Leo Grady. Random walks for image segmentation. *TPAMI*, 28(11):1768–1783, 2006.

Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *ICCV*, pp. 991–998, 2011.

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

Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas De Lange, Dag Johansen, and Håvard D Johansen. Kvasir-seg: A segmented polyp dataset. In *International conference on multimedia modeling*, pp. 451–462, 2019.

Lei Ke, Mingqiao Ye, Martin Danelljan, Yu-Wing Tai, Chi-Keung Tang, Fisher Yu, et al. Segment anything in high quality. *NeurIPS*, 36:29914–29934, 2023.

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. In *ICCV*, pp. 4015–4026, 2023.

Orest Kupyn, Hirokatsu Kataoka, and Christian Rupprecht. S3od: Towards generalizable salient object detection with synthetic data. In *ICLR*, 2026.

Trung-Nghia Le, Tam V Nguyen, Zhongliang Nie, Minh-Triet Tran, and Akihiro Sugimoto. Anabranch network for camouflaged object segmentation. *Computer vision and image understanding*, 184:45–56, 2019.

Feng Li, Qing Jiang, Hao Zhang, Tianhe Ren, Shilong Liu, Xueyan Zou, Huaizhe Xu, Hongyang Li, Jianwei Yang, Chunyuan Li, et al. Visual in-context prompting. In *CVPR*, pp. 12861–12871, 2024.

Yin Li, Xiaodi Hou, Christof Koch, James M Rehg, and Alan L Yuille. The secrets of salient object segmentation. In *CVPR*, pp. 280–287, 2014.

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*, pp. 740–755, 2014.

Qin Liu, Zhenlin Xu, Gedas Bertasius, and Marc Niethammer. Simpleclick: Interactive image segmentation with simple vision transformers. In *ICCV*, pp. 22290–22300, 2023.

Tie Liu, Zejian Yuan, Jian Sun, Jingdong Wang, Nanning Zheng, Xiaoou Tang, and Heung-Yeung Shum. Learning to detect a salient object. *TPAMI*, 33(2):353–367, 2010.

Xiangde Luo, Guotai Wang, Tao Song, Jingyang Zhang, Michael Aertsen, Jan Deprest, Sebastien Ourselin, Tom Vercauteren, and Shaoting Zhang. Mideepseg: Minimally interactive segmentation of unseen objects from medical images using deep learning. *Medical image analysis*, 72:102102, 2021.

Chaofan Ma, Yang Yuhuan, Chen Ju, Fei Zhang, Ya Zhang, and Yanfeng Wang. Attrseg: open-vocabulary semantic segmentation via attribute decomposition-aggregation. *NeurIPS*, 36:10258–10270, 2023.

Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15:654, 2024.

Kevis-Kokitsi Maninis, Sergi Caelles, Jordi Pont-Tuset, and Luc Van Gool. Deep extreme cut: From extreme points to object segmentation. In *CVPR*, pp. 616–625, 2018.

Zdravko Marinov, Paul F. Jäger, Jan Egger, Jens Kleesiek, and Rainer Stiefelhagen. Deep interactive segmentation of medical images: A systematic review and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10998–11018, 2024.

Vida Movahedi and James H Elder. Design and perceptual validation of performance measures for salient object segmentation. In *CVPR Workshop*, pp. 49–56, 2010.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research Journal*, 2024.

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *CVPR*, pp. 3498–3505, 2012.

Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *CVPR*, pp. 724–732, 2016.

Xuebin Qin, Hang Dai, Xiaobin Hu, Deng-Ping Fan, Ling Shao, and Luc Van Gool. Highly accurate dichotomous image segmentation. In *ECCV*, pp. 38–56, 2022.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. In *ICLR*, volume 2025, pp. 28085–28128, 2025.

Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. " grabcut" interactive foreground extraction using iterated graph cuts. *ACM Transactions on Graphics*, 23(3):309–314, 2004.

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

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Konstantin Sofiiuk, Ilia Petrov, Olga Barinova, and Anton Konushin. f-brs: Rethinking backpropagating refinement for interactive segmentation. In *CVPR*, pp. 8623–8632, 2020.

Konstantin Sofiiuk, Ilia A Petrov, and Anton Konushin. Reviving iterative training with mask guidance for interactive segmentation. In *ICIP*, pp. 3141–3145, 2022.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. Caltech-ucsd birds-200-2011, 2025.

Guotai Wang, Maria A Zuluaga, Rosalind Pratt, Michael Aertsen, Tom Doel, Maria Klusmann, Anna L David, Jan Deprest, Tom Vercauteren, and Sébastien Ourselin. Slic-seg: A minimally interactive segmentation of the placenta from sparse and motion-corrupted fetal MRI in multiple views. *Medical image analysis*, 34:137–147, 2016.

Guotai Wang, Wenqi Li, Maria A Zuluaga, Rosalind Pratt, Premal A Patel, Michael Aertsen, Tom Doel, Anna L David, Jan Deprest, Sébastien Ourselin, et al. Interactive medical image segmentation using deep learning with image-specific fine tuning. *IEEE transactions on medical imaging*, 37(7):1562–1573, 2018.

Guotai Wang, Maria A. Zuluaga, Wenqi Li, Rosalind Pratt, Premal A. Patel, Michael Aertsen, Tom Doel, Anna L. David, Jan Deprest, Sébastien Ourselin, and Tom Vercauteren. Deepigeos: A deep interactive geodesic framework for medical image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(7):1559–1572, 2019.

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*, pp. 136–145, 2017.

Yangtao Wang, Xi Shen, Shell Xu Hu, Yuan Yuan, James L Crowley, and Dominique Vaufreydaz. Self-supervised transformers for unsupervised object discovery using normalized cut. In *CVPR*, pp. 14543–14553, 2022.

Yunyang Xiong, Bala Varadarajan, Lemeng Wu, Xiaoyu Xiang, Fanyi Xiao, Chenchen Zhu, Xiaoliang Dai, Dilin Wang, Fei Sun, Forrest Iandola, et al. Efficientsam: Leveraged masked image pretraining for efficient segment anything. In *CVPR*, pp. 16111–16121, 2024.

Ning Xu, Brian Price, Scott Cohen, Jimei Yang, and Thomas Huang. Deep interactive object selection. In *CVPR*, pp. 373–381, 2016.

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

Sicheng Yang, Hongqiu Wang, Zhaohu Xing, Sixiang Chen, and Lei Zhu. Segdino: An efficient design for medical and natural image segmentation with dino-v3. *arXiv preprint arXiv:2509.00833*, 2025.

Chaoning Zhang, Dongshen Han, Yu Qiao, Jung Uk Kim, Sung-Ho Bae, Seungkyu Lee, and Choong Seon Hong. Faster segment anything: Towards lightweight sam for mobile applications. *arXiv preprint arXiv:2306.14289*, 2023.

Jianping Zhang, Lei Yang, Seyed Mahmoud Sajjadi Mohammadabadi, and Feng Yan. A survey on self-supervised learning: Recent advances and open problems. *Neurocomputing*, pp. 131409, 2025.