Learning Accurate Segmentation Purely from Self-Supervision

Zuyao You[1,2], Zuxuan Wu[1,2⋆], and Yu-Gang Jiang[1,2⋆]

[1] Institute of Trustworthy Embodied AI, Fudan University
[2] Shanghai Key Laboratory of Multimodal Embodied AI
zyyou23@m.fudan.edu.cn, {zxwu,ygj}@fudan.edu.cn

Abstract. Accurately segmenting objects without any manual annotations remains one of the core challenges in computer vision. In this work, we introduce **Selfment**, a fully self-supervised framework that segments foreground objects directly from raw images without human labels, pretrained segmentation models, or any post-processing. Selfment first constructs patch-level affinity graphs from self-supervised features and applies NCut to obtain an initial coarse foreground-background separation. We then introduce **Iterative Patch Optimization (IPO)**, a feature-space refinement procedure that progressively enforces spatial coherence and semantic consistency through iterative patch clustering. The refined masks are subsequently used as supervisory signals to train a lightweight segmentation head with contrastive and region-consistency objectives, allowing the model to learn stable and transferable object representations. Despite its simplicity and complete absence of manual supervision, Selfment sets new state-of-the-art (SoTA) results across multiple benchmarks. It achieves substantial improvements on F_{max} over previous unsupervised saliency detection methods on ECSSD (+4.0%), HKUIS (+4.6%), and PASCAL-S (+5.7%). Moreover, without any additional fine-tuning, Selfment demonstrates remarkable zero-shot generalization to camouflaged object detection tasks (*e.g.*, .910 S_m on CHAMELEON and .792 $\mathcal{F}_\beta^\omega$ on CAMO), outperforming all existing unsupervised approaches and even rivaling the SoTA fully supervised methods. Codes and weights are available at: https://geshang777.github.io/Selfment/.

Keywords: Self-supervised Learning · Unsupervised Saliency Detection · Unsupervised Camouflaged Object Detection

1 Introduction

Object segmentation has long relied on dense, human-annotated masks [8, 9, 26]. While these annotations provide precise supervision, they are costly and time-consuming to collect, limiting scalability and relying heavily on the human inductive bias. To reduce annotation overhead, recent efforts [15, 16, 27, 52] have explored weakly supervised cues (*i.e.*, points, scribbles, or motion trajectories, *etc.*)

⋆ Corresponding author.

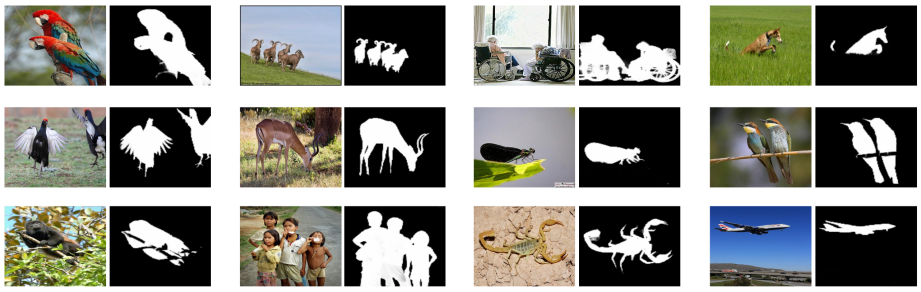

Fig. 1: We propose **Selfment**, a fully self-supervised framework for foreground segmentation that generates highly detailed and accurate saliency maps without any human-annotated labels or post-processing.

to guide segmentation. However, these methods still depend on human-provided signals and often rely on pretrained segmentation models (*e.g.*, SAM [18, 35]) through fine-tuning or prompt adaptation. As a result, they remain partially tied to manual supervision and pretrained segmentation representations.

This raises a fundamental question: Can a model learn accurate object segmentation directly from unlabeled images without any human annotations or external off-the-shelf segmentation models?

Recent advances in self-supervised learning (SSL) [2, 6, 13, 30, 43, 55] offer a promising path forward. SSL enables models to extract semantic structure directly from large-scale, unlabeled data. Among these methods, the DINO family [6, 30, 43] is particularly notable for producing strong, object-centric representations. Trained via a teacher-student self-distillation framework, DINO models align features across diverse image crops, encouraging semantically consistent regions to share similar embeddings. DINOv3 [43] further advances this line of work by introducing Gram Anchoring to stabilize dense patch features over long training schedules, enabling the model to scale to 7B parameters while maintaining high-quality representations.

The dense feature maps produced by DINO-series naturally encode semantic similarity, such that patches belonging to the same object or category tend to have highly similar embeddings. This property provides strong semantic priors for downstream tasks such as segmentation or object discovery. Building on this idea, several recent works [41, 48] treat the feature map as a graph and apply heuristic strategies, such as seed expansion or normalized graph cuts [38], to partition it into foreground and background. While these methods can generate approximate object masks, the resulting bipartitions are unstable, and the resulting masks tend to be coarse. Achieving reasonable segmentation quality typically requires heavy post-processing (*e.g.*, CRFs [19], bilateral solvers [3], or morphological refinements), which undermines the goal of the self-supervised segmentation.

In this work, we present **Selfment**, a fully self-supervised framework for foreground segmentation that requires no annotations, no post-processing, and no off-the-shelf segmentation models. Built on DINOv3 [43], our approach begins by constructing a patch-wise affinity graph based on the image feature and

applying Normalized Cut [38] to derive an initial bipartition, yielding a coarse yet semantically grounded segmentation prior. To further enhance spatial coherence and reduce noise from spectral relaxation, we introduce the **Iterative Patch Optimization (IPO)**, a simple yet effective module that refines assignments by iteratively clustering patches in the feature space. At each step, foreground and background centroids are updated, and patch labels are realigned according to semantic similarity. Orientation consistency constraints further stabilize the refinement, preventing degenerate solutions and ensuring that the evolving segmentation adheres to a coherent object-background separation. The refined masks then serve as self-supervised signals to train a lightweight segmentation head with **contrastive and region-consistency objectives**, enabling the model to progressively learn discriminative, object-aware representations.

Without any post-processing, Selfment surpasses the performance of previous state-of-the-art (SoTA) models by a very clear margin. On widely used salient object detection benchmarks such as ECSSD [39], DUTS [47], HKUIS [23], and PASCAL-S [25], Selfment yields substantial improvements of 4.0%, 7.0%, 4.6%, and 5.7% in F_{max}, respectively. As illustrated in Fig. 1, Selfment can generate accurate and highly detailed saliency maps with the input resolution of 2048×2048, purely based on our proposed self-supervised method. Furthermore, without any task-specific fine-tuning, Selfment demonstrates remarkable zero-shot generalization on camouflaged object detection tasks, achieving an S_m of .910, .869, .873, and .902 on CHAMELEON [44], CAMO [21], COD10K [11], and NC4K [28], outperforming all previous unsupervised approaches and approaching the performance of fully supervised methods.

Our contributions are summarized as follows:

- We introduce **Selfment**, a fully self-supervised segmentation framework that operates without human annotations, external segmentation priors, or post-processing steps.
- We introduce a simple yet effective mask refinement algorithm based on patch similarity, which significantly improves the performance of the initial NCut segmentation. Moreover, the method can be easily transferred across different self-supervised backbones.
- Extensive experiments demonstrate that Selfment establishes new state-of-the-art results on both unsupervised salient object detection and camouflaged object detection tasks.

2 Related Work

2.1 Self-supervised Vision Foundation Models

Self-supervised vision foundation models [6, 13, 22, 30, 43] have advanced rapidly as backbone models trained without human annotations, enabling strong transfer to dense prediction tasks [8, 41, 48, 51]. Early work [7, 14, 33] explored contrastive objectives, which established that invariances learned from augmentations can

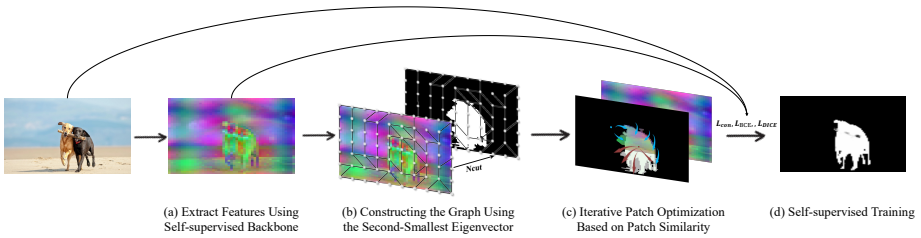

Fig. 2: An overview of Selfment. The input image is first encoded by a self-supervised backbone to produce dense patch features. These features define a patch-level affinity graph, from which we derive an initial foreground-background split using the second-smallest eigenvector of the NCut. We then apply Iterative Patch Optimization to improve spatial coherence and semantic consistency. The refined masks then serve as supervisory signals for training a lightweight segmentation head.

yield competitive semantic features. Masked image modeling introduced an alternative paradigm [1, 13], demonstrating that reconstructing masked patches produces scalable and robust backbone models. More recent foundation models [6, 30, 43] have focused on dense, spatially consistent representations, with the latest DINOv3 [43] further strengthening dense feature stability through long-schedule training and Gram anchoring. Our work builds on this line by leveraging dense representations from DINOv3 and developing a fully self-supervised framework that converts these features into reliable foreground segmentation.

2.2 Unsupervised Object Segmentation

Unsupervised object segmentation aims to separate foreground objects from background without relying on any manual annotations. Early unsupervised object segmentation methods relied heavily on low-level cues such as color contrast, texture consistency, motion boundaries, or super-pixel grouping [24, 27, 52, 56], which limited their robustness in complex scenes. More recent approaches alleviate these limitations by leveraging powerful off-the-shelf segmentation models such as SAM [18] to provide pseudo labels or guidance [15, 16], yet this dependence introduces strong external priors and reduces the level of true unsupervision. TokenCut [48] takes a different direction by using self-supervised ViT features and applying a normalized-cut formulation [38] to patch similarities, showing that object regions can emerge directly from transformer attention. However, its bipartitions are often unstable, and achieving competitive segmentation quality typically requires heavy post-processing, including bilateral solvers [3] or CRFs [19]. In contrast, our approach builds segmentation directly from self-supervised dense features, requires no annotated masks, does not rely on any external segmentation model such as SAM [18, 35], and achieves high-quality foreground segmentation without any post-processing.

3 Method

3.1 Normalized Cut (NCut)

Given an image represented by a set of patch features $\{f_i\}_{i=1}^{N}$ extracted from the pretrained backbone, we construct an undirected weighted graph $G = (V, E)$, where each node $v_i \in V$ corresponds to a patch, and each edge $e_{ij} \in E$ represents the similarity between patches i and j. The pairwise affinity is defined as:

$$A_{ij} = \begin{cases} \langle f_i, f_j \rangle, & \text{if } \langle f_i, f_j \rangle > \tau, \\ \epsilon, & \text{otherwise,} \end{cases} \tag{1}$$

where τ is a similarity threshold set to 0.2 and ϵ is a small constant to ensure graph connectivity. Let D denote the diagonal degree matrix with entries $D_{ii} = \sum_j A_{ij}$.

Following [38, 48], the NCut objective seeks to partition the graph into two disjoint sets A and B such that the inter-group similarity is minimized while maintaining high intra-group affinity:

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

where $\text{cut}(X, Y) = \sum_{i \in X, j \in Y} X_{ij}$ measures the degree of similarity between two sets. Minimizing $\text{NCut}(A, B)$ is equivalent to solving the generalized eigenvalue problem:

$$(D - A)\mathbf{x} = \lambda D\mathbf{x}, \tag{3}$$

where $\mathbf{x}$ is a relaxed continuous indicator vector representing the partition. The smallest eigenvector corresponds to the trivial constant solution, while the second smallest eigenvector $\mathbf{x}_2$ (also known as the *Fiedler vector*) defines the optimal bipartition of the graph.

The binary segmentation mask is obtained by thresholding $\mathbf{x}_2$ at its mean value:

$$\text{mask}(i) = \begin{cases} 1, & \text{if } x_{2,i} > \frac{1}{N} \sum_{j=1}^{N} x_{2,j}, \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

Finally, to identify the principal object, we locate the connected component that contains the seed patch (corresponding to the maximum absolute value in $\mathbf{x}_2$), and treat it as the main object mask. This procedure yields a coarse foreground-background bipartition, serving as the initialization for our subsequent refinement stage.

3.2 Iterative Patch Optimization

The bipartition obtained from the normalized cut (Sec. 3.1) is often noisy and spatially inconsistent due to the binary graph construction and the spectral relaxation. To address this, we design an iterative refinement procedure that exploits the semantic similarity of patch features derived from self-supervised representations to progressively improve patch-level consistency.

Given the feature map $F = \{f_i \in \mathbb{R}^d\}_{i=1}^N$, we first normalize all patch embeddings by:

$$\tilde{f}_i = \frac{f_i}{\|f_i\|_2}. \tag{5}$$

Let $\mathbf{y}^{(0)} \in \{0,1\}^N$ denote the initial bipartition mask obtained from the NCut result, where 1 indicates foreground. We compute initial cluster centroids for the foreground and background regions as

$$\mu_f^{(0)} = \frac{1}{|\mathcal{F}^{(0)}|} \sum_{i \in \mathcal{F}^{(0)}} \tilde{f}_i, \quad \mu_b^{(0)} = \frac{1}{|\mathcal{B}^{(0)}|} \sum_{i \in \mathcal{B}^{(0)}} \tilde{f}_i, \tag{6}$$

where $\mathcal{F}^{(0)} = \{i \mid y_i^{(0)} = 1\}$ and $\mathcal{B}^{(0)} = \{i \mid y_i^{(0)} = 0\}$.

At each iteration t, we update the label of every patch according to its relative similarity to the current cluster means:

$$y_i^{(t+1)} = \begin{cases} 1, & \text{if } \langle \tilde{f}_i, \mu_f^{(t)} \rangle > \langle \tilde{f}_i, \mu_b^{(t)} \rangle, \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

After relabeling, the new cluster means are recomputed as:

$$\begin{aligned} \mu_f^{(t+1)} &= \frac{1}{|\mathcal{F}^{(t+1)}|} \sum_{i \in \mathcal{F}^{(t+1)}} \tilde{f}_i, \\ \mu_b^{(t+1)} &= \frac{1}{|\mathcal{B}^{(t+1)}|} \sum_{i \in \mathcal{B}^{(t+1)}} \tilde{f}_i. \end{aligned} \tag{8}$$

The process is repeated for a fixed number of iterations $T = 20$. To avoid label flipping between iterations, we enforce orientation consistency by maintaining a reference vector $\mathbf{r} = \mu_f^{(0)} - \mu_b^{(0)}$, and reversing labels when $\langle (\mu_f^{(t+1)} - \mu_b^{(t+1)}), \mathbf{r} \rangle < 0$.

This iterative optimization refines the NCut initialization by aligning patch assignments with semantically similar features learned from self-supervised training. Our refinement relies solely on feature similarity, producing significantly cleaner and semantically coherent masks without any external priors or annotations.

3.3 Self-supervised Training

As illustrated in Fig. 2, we employ masks from Sec. 3.2 as self-supervised signals to train a lightweight segmentation head that learns discriminative patch embeddings for more robust and stable unsupervised salient object detection.

We introduce a two-layer projection head followed by a binary classifier, denoted as ϕ_θ, that operates on patch features $f_i \in \mathbb{R}^d$ extracted from the vision backbone. The projection head maps features into an embedding space via:

$$z_i = \phi_\theta(f_i) = W_2 \sigma(W_1 f_i + b_1) + b_2, \tag{9}$$

where $\sigma(\cdot)$ is a ReLU activation. The classifier then outputs logits $l_i = W_c z_i + b_c \in \mathbb{R}^2$ corresponding to foreground and background probabilities.

For each image, we obtain patch-level pseudo-labels $\mathbf{y} \in \{0,1\}^N$ from the iterative patch optimization described in Sec. 3.2. These pseudo-labels provide noisy but spatially consistent foreground supervision that guides the patch embedding learning.

Inspired by InfoNCE [36], we impose a feature-level alignment encouraging embeddings of patches from the same region (foreground or background) to be close while pushing apart those from opposite regions. Let z_i be ℓ_2-normalized patch embeddings, and $S = z_i^\top z_j/\tau$ be their pairwise similarity with temperature τ. We define:

$$\mathcal{L}_{\text{con}} = -\frac{1}{K}\sum_{i=1}^{K}\frac{1}{|\mathcal{P}_i|}\sum_{j\in\mathcal{P}_i}\log\frac{\exp(S_{ij})}{\sum_{k\neq i}\exp(S_{ik})}, \tag{10}$$

where $\mathcal{P}_i = \{j \neq i \mid y_j = y_i\}$ denotes positive pairs sharing the same label. To further encourage segmentation consistency, we use the soft Dice loss:

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2\sum_i p_i y_i + \epsilon}{\sum_i p_i^2 + \sum_i y_i^2 + \epsilon}, \tag{11}$$

where $p_i = \sigma(l_i^{(1)})$ are the predicted foreground probabilities. Besides, each patch is trained to predict its pseudo-label via BCE loss:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}\Big[y_i \log\sigma(l_i^{(1)}) + (1-y_i)\log(1-\sigma(l_i^{(1)}))\Big], \tag{12}$$

where $\sigma(\cdot)$ denotes the sigmoid and $l_i^{(1)}$ is the foreground logit. The overall self-supervised loss is a weighted combination of these components:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{con}}\mathcal{L}_{\text{con}} + \lambda_{\text{Dice}}\mathcal{L}_{\text{Dice}} + \lambda_{\text{BCE}}\mathcal{L}_{\text{BCE}}, \tag{13}$$

where λ_{con}, λ_{Dice}, and λ_{BCE} are set to 0.1, 1.0, and 1.0, respectively. This objective encourages the model to learn discriminative, spatially coherent patch embeddings that capture objectness purely from self-supervised cues without any labeled data.

4 Experiment

4.1 Implementation Details

All experiments are implemented with DINOv3-7B and kept frozen during training. We randomly sample 1,000 images from the DUTS [47] training set and use these as our self-supervised training corpus. The images are resized to the resolution of 768×768 for the training stages. The patch-head (Sec. 3.3) is optimized with Adam at a learning rate of 1×10^{-3} for 3 epochs. To accelerate training, we extract and cache backbone patch features before head training and load these cached features from fast storage during each epoch. All experiments are carried out using distributed training on 8 NVIDIA A100 GPUs with 80G memory with PyTorch DistributedDataParallel [32]. For reproducibility, we report averaged results where applicable.

Table 1: Comparison of unsupervised saliency detection methods. We evaluate Selfment against state-of-the-art unsupervised methods on ECSSD [39], DUTS [47], HKUIS [23], and PASCAL-S [25]. Best results are **bolded**.

	ECSSD [39]			DUTS [47]			HKUIS [23]			PASCAL-S [25]		
	$F_{\max}$	IoU	Acc.	$F_{\max}$	IoU	Acc.	$F_{\max}$	IoU	Acc.	$F_{\max}$	IoU	Acc.
HS [49]	67.3	50.8	84.7	50.4	36.9	82.6	-	-	-	-	-	-
wCtr [56]	68.4	51.7	86.2	52.2	39.2	83.5	-	-	-	-	-	-
WSC [24]	68.3	49.8	85.2	52.8	38.4	86.2	-	-	-	-	-	-
DeepUSPS [29]	58.4	44.0	79.5	42.5	30.5	77.3	-	-	-	-	-	-
BigBiGAN [46]	78.2	67.2	89.9	60.8	49.8	87.8	-	-	-	-	-	-
E-BigBiGAN [46]	79.7	68.4	90.6	62.4	51.1	88.2	-	-	-	-	-	-
LOST [37, 41]	75.8	65.4	89.5	61.1	51.8	87.1	-	-	-	-	-	-
TokenCut-768 × 768 [48]	87.8	75.9	92.9	73.0	60.5	90.3	82.2	64.0	92.1	76.2	61.3	84.0
TokenCut-1280 × 1280 [48]	86.7	75.0	92.3	68.1	55.9	87.9	79.6	64.9	92.6	80.0	60.2	83.2
SelfMask-768 × 768 [40]	91.9	77.2	93.9	79.4	62.4	92.4	89.8	74.4	94.7	86.0	61.8	85.5
SelfMask-1280 × 1280 [40]	89.9	72.1	92.1	68.1	55.9	87.9	88.2	69.6	93.4	84.8	58.5	84.1
FOUND-768 × 768 [42]	91.4	78.9	94.4	76.4	64.8	93.7	87.7	68.4	93.8	85.4	63.7	86.3
FOUND-1280 × 1280 [42]	89.2	71.9	92.5	72.6	55.1	91.5	86.6	64.5	92.9	84.0	60.0	84.9
Selfment (Ours)-768 × 768	95.3	82.4	95.2	85.1	66.6	91.8	93.9	80.0	95.8	91.5	71.2	88.8
Selfment (Ours)-1280 × 1280	**95.9**	**84.3**	**95.8**	**86.4**	**68.4**	**92.3**	**94.4**	**81.6**	**96.1**	**91.7**	**71.6**	**88.8**

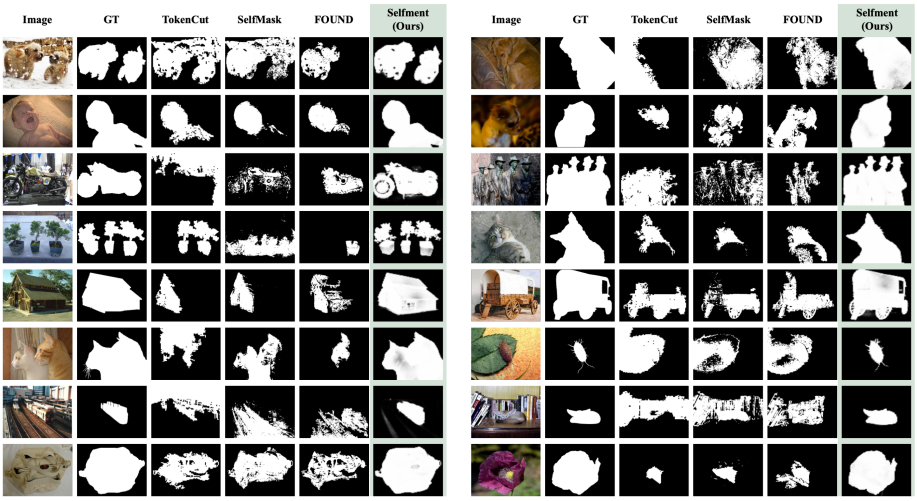

Fig. 3: Comparison with previous state-of-the-art methods on the unsupervised saliency detection task. All methods are evaluated without any post-processing at an inference resolution of 1280 × 1280.

4.2 Benchmarks and Evaluation Metrics

Unsupervised saliency detection. Unsupervised saliency detection aims to identify and segment the most salient object in an image without supervision. We compare our method with a wide range of unsupervised saliency detection

Table 2: Comparison of camouflaged object detection methods. We evaluate Selfment against state-of-the-art methods on CHAMELEON [44], CAMO [21], COD10K [11], and NC4K [28]. Best results are **bolded**.

	CHAMELEON [44]				CAMO [21]				COD10K [11]				NC4K [28]			
	$\mathcal{S}_m \uparrow$	$\mathcal{F}_\beta^\omega \uparrow$	$E_\xi \uparrow$	MAE	$\mathcal{S}_m \uparrow$	$\mathcal{F}_\beta^\omega \uparrow$	$E_\xi \uparrow$	MAE	$\mathcal{S}_m \uparrow$	$\mathcal{F}_\beta^\omega \uparrow$	$E_\xi \uparrow$	MAE	$\mathcal{S}_m \uparrow$	$\mathcal{F}_\beta^\omega \uparrow$	$E_\xi \uparrow$	MAE
Fully-Supervised Methods																
BGNet [45]	.901	.851	.954	.027	.812	.749	.870	.073	.831	.722	.901	.033	.851	.788	.907	.044
SINetv2 [10]	.888	.816	.961	.030	.820	.743	.882	.070	.815	.680	.887	.037	.847	.770	.903	.048
ZoomNet [31]	.902	.845	.958	.023	.820	.752	.878	.066	.838	.729	.888	.029	.853	.784	.896	.043
FSPNet [17]	.908	.851	.965	.023	.856	.799	.899	.050	.851	.735	.895	.026	.879	.816	.915	.035
BiRefNet [54]	.929	.911	.968	.016	.932	.914	.974	.015	.913	.874	.960	.014	.914	.894	.953	.023
Semi-Supervised Methods																
CamoTeacher [20]	.756	.617	.813	.065	.701	.560	.795	.112	.759	.594	.854	.049	.791	.687	.868	.068
SCOD-ND [12]	.850	.773	.928	.036	.789	.732	.859	.077	.819	.725	.891	.033	.838	.787	.903	.046
Unsupervised Methods																
BigBiGAN [46]	.547	.244	.527	.257	.565	.299	.528	.282	.528	.185	.497	.261	.608	.319	.565	.246
TokenCut [48]	.654	.496	.740	.132	.633	.498	.706	.163	.658	.469	.735	.103	.725	.615	.802	.101
SelfMask [40]	.619	.436	.675	.176	.617	.483	.698	.176	.637	.431	.679	.131	.716	.593	.777	.114
UCOS-DA [53]	.750	.639	.808	.091	.702	.604	.751	.148	.655	.467	.687	.120	.731	.617	.785	.103
UCOD-DPL [50]	.864	.825	.931	.031	.793	.747	.862	.077	.834	**.763**	**.916**	**.031**	.850	.818	.923	.043
Selfment (Ours)	**.910**	**.843**	**.944**	**.025**	**.869**	**.792**	**.894**	**.060**	**.873**	.754	.900	.033	**.902**	**.836**	**.931**	**.033**

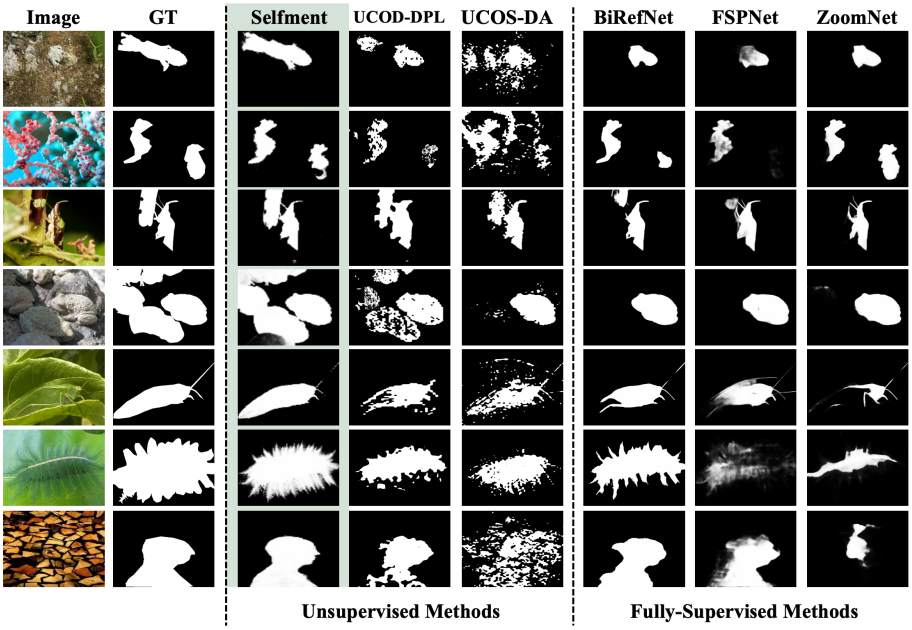

Fig. 4: Comparison with previous SoTA on the camouflaged object detection tasks.

approaches. Table 1 reports the quantitative results on ECSSD [39], DUTS [47], HKUIS [23], and PASCAL-S [25]. All the methods are compared in terms of F_{max}, IoU, and pixel accuracy (Acc.) following previous work [40, 42, 48].

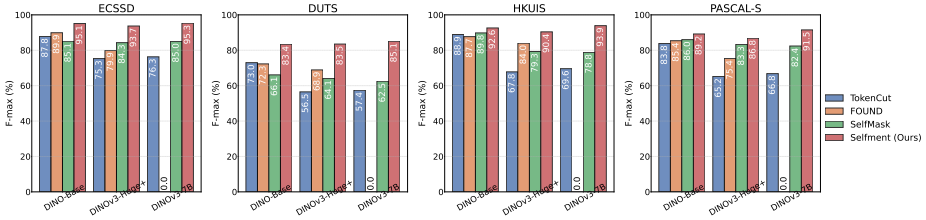

Fig. 5: Comparison of segmentation performance among TokenCut [48], SelfMask [40], FOUND [42], and Selfment using DINO-Base, DINOv3-Huge+, and DINOv3-7B as backbones. Metrics are reported on the ECSSD [39] dataset.

Camouflaged object detection (COD). Camouflaged object detection is an extremely challenging segmentation task that requires identifying objects seamlessly concealed within their surroundings. We evaluate Selfment on CAMO [21], COD10K [11], NC4K [28], and CHAMELEON [44], which are commonly use in COD tasks. We evaluate segmentation accuracy using S-Measure ($\mathcal{S}_m$), E-Measure (E_ξ), Weighted F-Measure ($\mathcal{F}_\beta^\omega$), and MAE following previous work [10, 31, 54].

4.3 Main Results

Results on unsupervised saliency detection. The works most similar to ours are [40–42, 48], which also rely on feature embeddings from self-supervised foundations. For fair comparison, we standardize the input resolution during inference and compare these methods with Selfment at the resolution of 768×768 and 1280×1280. As shown in the Tab. 1, Selfment outperforms these approaches across all metrics, achieving substantial improvements. Additionally, Selfment consistently benefits from increasing input image resolution during the inference stage. In contrast, other models tend to experience a decline in performance as the resolution increases. We further discuss this behavior in the Sec. 4.4.

We also present a qualitative comparison between Selfment and previous methods in Fig. 3. Benefiting from the iterative patch optimization and robust self-supervised training, Selfment produces saliency maps that are both complete and precise. In contrast, existing unsupervised methods tend to generate fragmented and incomplete predictions.

***Zero-shot* on COD.** We evaluate the zero-shot performance of Selfment on COD. Remarkably, Selfment not only surpasses all previous unsupervised methods by a large margin but also outperforms several fully supervised methods. For example, as shown in the Tab. 2, Selfment achieves .869 on $\mathcal{S}_m$ on CAMO [21], significantly outperforming the previous unsupervised method (+.076), and even surpassing strong fully supervised approaches such as FSPNet [17]. Fig. 4 further provides qualitative comparisons. Without any task-specific fine-tuning or post-processing, Selfment accurately detects camouflaged objects and produces detailed, high-quality saliency maps. These results demonstrate the strong generalization ability

of Selfment and highlight the potential of self-supervised learning for challenging segmentation tasks such as COD.

4.4 Ablation Study

We conduct ablation studies on the ECSSD [39] dataset to investigate the effect of our proposed modules and methods. For the experiments reported in Table 3, we use a subset of 500 images sampled from DUTS for self-supervised training to allow agile iteration, fix the data sampling seed while keeping all other settings identical to those described in Sec. 4.1. The remaining experiments follow the same recipe as reported in Table 1. All experiments are conducted with strict variable control.

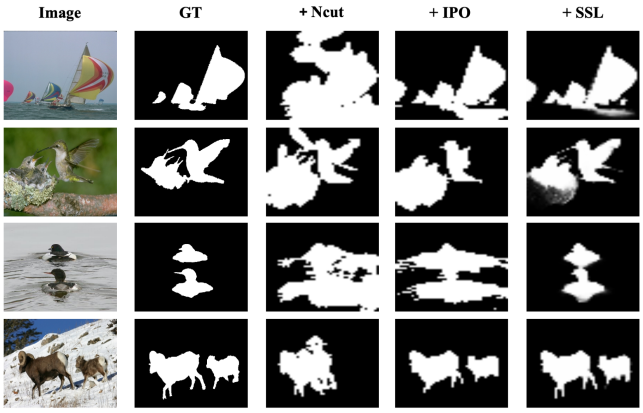

Fig. 6: Ablation study on the main components of the Selfment pipeline. From left to right, we progressively add NCut, IPO, and self-supervised learning (SSL).

Ablation on Backbone In this section, we extensively compare the performance of Tokencut [48], SelfMask [40], FOUND [42], and Selfment under the same backbone and resolution to explore the impact of different self-supervised foundation models on model performance. We use DINO-Base, DINOv3-Huge+, and DINOv3-7B [43] as backbones. Notably, for SelfMask, due to the large size of the model, it cannot be trained on a single A100 GPU, so we used DeepSpeed Zero3 [34] for model sharding, while keeping other configurations unchanged. In addition, because the feature maps produced by DINO-Huge+ do not exhibit reliable bipartition properties under NCut, both TokenCut and Selfment use the value features from the last self-attention layer when DINO-Huge+ is adopted as the backbone. All experiments are conducted with a fixed inference resolution of 768×768.

As shown in the Fig. 5, Selfment consistently outperforms previous methods by a clear margin, regardless of the backbone used. It is worth noting that, except for Selfment, the other models do not benefit from model scaling. In particular, FOUND exhibits instability during training when using DINOv3-7B as the backbone, resulting in both F-max being zero. This is because FOUND heavily relies on the initial background seed for similarity computation in the initialization step. However, due to the semantics of DINOv3 being very fine-grained, the similarity computed solely from the background seed leads to very

Table 3: Ablation study of individual components. Each module is added cumulatively to the previous configuration. BCE: binary cross-entropy loss; Dice: soft Dice loss; Con.: contrastive similarity loss.

Configuration	BCE	Dice	Con.	F_{max}	IoU	Acc
NCut (baseline)				74.7	63.9	86.2
+ IPO				79.5	73.2	87.8
Self-supervised training	✓			88.3	80.4	94.4
+ Dice loss	✓	✓		88.9	81.3	94.7
+ Contrastive loss	✓		✓	88.9	81.4	94.7
+ Dice & Contrastive loss	✓	✓	✓	**89.1**	**81.5**	**94.8**

fragmented pseudo-labels, making it difficult for the model to learn accurate semantics from the pseudo-labels, ultimately causing the training to fail. In contrast, Selfment effectively leverages the semantic structure encoded in both DINO and DINOv3 through components such as iterative patch optimization, enabling it to robustly exploit patch-level similarity and semantic consistency across different self-supervised backbones. Consequently, Selfment can achieve strong and stable performance regardless of backbone choice.

Effect of IPO. IPO leverages feature similarity in the DINOv3 embedding space to iteratively refine patch assignments, producing more coherent and semantically meaningful object masks without any additional supervision. As shown in Fig. 7, IPO substantially improves the ambiguous masks produced by the initial NCut step, quickly converging within roughly 10 iterations to a mask that closely aligns with the object boundary while preserving fine-grained details. Quantitatively, IPO yields a significant performance gain (Table 3), improving F_{max}, IOU, and accuracy by 4.8%, 9.3%, and 1.6%, respectively.

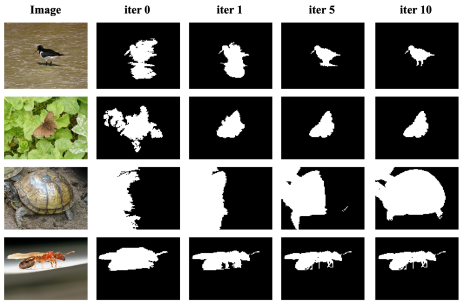

Fig. 7: Progressive refinement of the mask across iterations of the IPO procedure.

Self-supervised training. We further employ the pseudo-labels generated by IPO to train a lightweight patch-level head in a self-supervised manner, enhancing the robustness and stability of the predicted saliency maps. As shown in Table 3, training with these pseudo-labels significantly boosts segmentation accuracy. Even when using only the BCE loss, the model reaches an F_{max} of 88.3%, a substantial improvement over the 79.5% obtained without the self-supervised training stage, indicating that the learned embeddings can effectively extract objectness cues from noisy pseudo-labels alone. Building on the BCE loss, we further incorporate

the contrastive similarity loss and the Dice loss. The contrastive similarity loss encourages the model to exploit patch-level feature relationships by pulling together patches belonging to the same region while pushing apart those from different regions. Incorporating this loss leads to an additional performance gain, improving F_{max} from 88.3% to 88.9% and IoU from 80.4% to 81.4%. In addition, adding the Dice loss promotes spatial compactness and boundary completeness, contributing a further +0.2% improvement in IoU.

Effect of Input Resolution. In this section, we analyze how Selfment and TokenCut [48] behave as the input resolution increases in Fig. 8. For a fair comparison, both methods operate on the output feature from the last layer of the DINOv3-7B. We evaluate saliency map quality by resizing the input images to three resolutions: 768×768, 1536×1536, and 2560×2560. Although Selfment is trained only at 768×768 resolution, it generalizes naturally to much higher resolutions. As the input size increases, the predicted saliency maps become progressively sharper and more detailed, while maintaining object-level coherence, *e.g.*, the text on the signboard in the last row. TokenCut, however, depends heavily on a single NCut bipartition, which becomes unstable on high-resolution affinity graphs. As the resolution grows, this instability leads to noticeable degradation in its predictions.

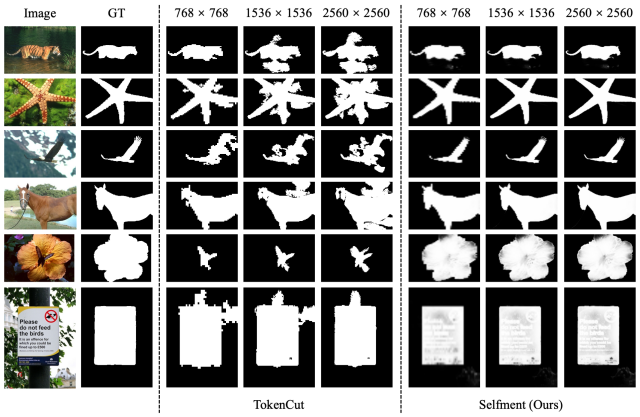

Fig. 8: Qualitative comparison of how Selfment and TokenCut behave as input resolution increases.

Comparison of Initial Bipartition Methods We compare different initial bipartition strategies in this section: using the DINOv3 `<CLS>` token, K-means, and NCut [38], and report their performance on ECSSD [39]. As shown in Tab. 4, using the `<CLS>` token to derive a binary partition performs poorly. The resulting bipartition yields only an F_{max} of 29.2 on ECSSD. A simple K-means clustering over patch embeddings also offers a reasonable bipartition, but its quality remains significantly worse than that of NCut. In contrast, NCut provides a substantially more reliable initialization. Leveraging the second smallest eigenvector to guide the partition produces a

Table 4: Comparison of initial bipartition methods on the ECSSD.

	F_{max}	IoU	Acc
`<CLS>`	29.2	20.5	54.0
K-Means	59.2	55.1	71.7
NCut	74.7	63.9	86.2

meaningful foreground–background separation, achieving 74.7 $F_{\max}$, 63.9 IoU, and 86.2 accuracy.

Generalization beyond DINO. To assess whether the effectiveness of Selfment extends beyond DINO-based representations, we conduct ablation studies using two alternative ViT backbones: Perception Encoder (PE) [4], trained with contrastive learning, and TIPSv2 [5], trained with a hybrid objective combining contrastive and self-supervised learning. As shown in the Tab. 5, Selfment consistently benefits from both IPO and self-training across all backbone variants, demonstrating the robustness of our proposed framework.

Table 5: Comparison with TokenCut under PE-Spacial and TIPSv2.

Backbone	Method	$F_{\max}$ ↑	IoU ↑	Acc. ↑
PE-Spacial B/16 [4]	TokenCut	56.4	53.8	80.0
	Selfment	**92.6**	**68.6**	**89.9**
PE-Spacial G/14 [4]	TokenCut	58.3	55.4	80.9
	Selfment	**91.7**	**75.1**	**93.1**
TIPSv2 B/14 [5]	TokenCut	51.5	46.6	74.4
	Selfment	**88.7**	**64.6**	**87.8**
TIPSv2 SO400M/14 [5]	TokenCut	63.4	56.0	82.5
	Selfment	**90.5**	**69.1**	**90.9**

4.5 Computational Efficiency

During training, we resize all images to 768 × 768. Unless otherwise specified, all experiments adopt the DINOv3-7B as backbone. The backbone remains frozen throughout training, and only the lightweight segmentation head described in Sec. 3.3 is optimized. This head contains merely 0.54M trainable parameters and requires 1.08M FLOPs per forward pass. Furthermore, we cache the backbone features to avoid redundant computation. Using 8×A100 GPUs, Selfment is trained for 3 epochs in only 27.6 minutes. As shown in Tab. 6, during inference, a single A100 GPU processes an 768 × 768 image in 0.029s with cached features and 1.915s without caching, which remains practical given the scale of the DINOv3-7B.

Table 6: Inference time per image (seconds) for TokenCut and Selfment under different backbones, with and without feature caching.

Method	Backbone	Cache	Time (s)
Tokencut	DINOv3 7B/16	×	5.087
	DINO B/16	×	0.771
Selfment	DINOv3 7B/16	✓	**0.029**
	DINOv3 7B/16	×	1.915
	DINO B/16	✓	**0.003**
	DINO B/16	×	0.165

5 Conclusion

In this work, we presented Selfment, a fully self-supervised segmentation framework that achieves state-of-the-art performance without relying on human an-

notations, post-processing, or priors from SAM. By leveraging self-supervised feature learning and an iterative patch optimization method, Selfment produces highly accurate saliency maps and outperforms existing approaches on multiple benchmarks. Additionally, it can be easily transferred to camouflage object detection tasks, significantly surpassing previous methods. Selfment demonstrates that high-quality object segmentation can be achieved entirely through self-supervision, setting a new standard for fully autonomous, annotation-free segmentation.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (Grant No. 62427819) and the Science and Technology Commission of Shanghai Municipality (No. 25511106100).

References

1. Bao, H., Dong, L., Piao, S., Wei, F.: Beit: Bert pre-training of image transformers. arXiv preprint arXiv:2106.08254 (2021)
2. Bardes, A., Garrido, Q., Ponce, J., Chen, X., Rabbat, M., LeCun, Y., Assran, M., Ballas, N.: Revisiting feature prediction for learning visual representations from video. arXiv preprint arXiv:2404.08471 (2024)
3. Barron, J.T., Poole, B.: The fast bilateral solver. In: ECCV (2016)
4. Bolya, D., Huang, P.Y., Sun, P., Cho, J.H., Madotto, A., Wei, C., Ma, T., Zhi, J., Rajasegaran, J., Bangalath, H., et al.: Perception encoder: The best visual embeddings are not at the output of the network. NeurIPS (2026)
5. Cao, B., Chen, K., Maninis, K.K., Chen, K., Karpur, A., Xia, Y., Dua, S., Dabral, T., Han, G., Han, B., et al.: Tipsv2: Advancing vision-language pretraining with enhanced patch-text alignment. In: CVPR (2026)
6. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: ICCV (2021)
7. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: ICML (2020)
8. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: CVPR (2022)
9. Cheng, B., Schwing, A., Kirillov, A.: Per-pixel classification is not all you need for semantic segmentation. NeurIPS (2021)
10. Fan, D.P., Ji, G.P., Cheng, M.M., Shao, L.: Concealed object detection. TPAMI (2021)
11. Fan, D.P., Ji, G.P., Sun, G., Cheng, M.M., Shen, J., Shao, L.: Camouflaged object detection. In: CVPR (2020)
12. Fu, Y., Ying, J., Lv, H., Guo, X.: Semi-supervised camouflaged object detection from noisy data. In: ACM MM (2024)
13. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: CVPR (2022)
14. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: CVPR (2020)
15. Hu, J., Cheng, Z., Gong, S.: Int: Instance-specific negative mining for task-generic promptable segmentation. arXiv preprint arXiv:2501.18753 (2025)

16. Hu, J., Lin, J., Gong, S., Cai, W.: Relax image-specific prompt requirement in sam: A single generic prompt for segmenting camouflaged objects. In: AAAI (2024)
17. Huang, Z., Dai, H., Xiang, T.Z., Wang, S., Chen, H.X., Qin, J., Xiong, H.: Feature shrinkage pyramid for camouflaged object detection with transformers. In: CVPR (2023)
18. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: ICCV (2023)
19. Krähenbühl, P., Koltun, V.: Efficient inference in fully connected crfs with gaussian edge potentials. NeurIPS (2011)
20. Lai, X., Yang, Z., Hu, J., Zhang, S., Cao, L., Jiang, G., Wang, Z., Zhang, S., Ji, R.: Camoteacher: Dual-rotation consistency learning for semi-supervised camouflaged object detection. In: ECCV (2024)
21. Le, T.N., Nguyen, T.V., Nie, Z., Tran, M.T., Sugimoto, A.: Anabranch network for camouflaged object segmentation. CVIU (2019)
22. LeCun, Y.: A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. Open Review (2022)
23. Li, G., Yu, Y.: Visual saliency based on multiscale deep features. In: CVPR (2015)
24. Li, N., Sun, B., Yu, J.: A weighted sparse coding framework for saliency detection. In: CVPR (2015)
25. Li, Y., Hou, X., Koch, C., Rehg, J.M., Yuille, A.L.: The secrets of salient object segmentation. In: CVPR (2014)
26. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV (2014)
27. Liu, X., Huang, X.: Weakly supervised salient object detection via bounding-box annotation and sam model. ERA (2024)
28. Lv, Y., Zhang, J., Dai, Y., Li, A., Liu, B., Barnes, N., Fan, D.P.: Simultaneously localize, segment and rank the camouflaged objects. In: CVPR (2021)
29. Nguyen, T., Dax, M., Mummadi, C.K., Ngo, N., Nguyen, T.H.P., Lou, Z., Brox, T.: Deepusps: Deep robust unsupervised saliency prediction via self-supervision. NeurIPS (2019)
30. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Howes, R., Huang, P.Y., Xu, H., Sharma, V., Li, S.W., Galuba, W., Rabbat, M., Assran, M., Ballas, N., Synnaeve, G., Misra, I., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision (2023)
31. Pang, Y., Zhao, X., Xiang, T.Z., Zhang, L., Lu, H.: Zoom in and out: A mixed-scale triplet network for camouflaged object detection. In: CVPR (2022)
32. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. NeurIPS (2019)
33. Purushwalkam, S., Gupta, A.: Demystifying contrastive self-supervised learning: Invariances, augmentations and dataset biases. NeurIPS (2020)
34. Rasley, J., Rajbhandari, S., Ruwase, O., He, Y.: Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In: KDD (2020)
35. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
36. Rusak, E., Reizinger, P., Juhos, A., Bringmann, O., Zimmermann, R.S., Brendel, W.: Infonce: Identifying the gap between theory and practice. arXiv preprint arXiv:2407.00143 (2024)

37. Shen, X., Efros, A.A., Joulin, A., Aubry, M.: Learning co-segmentation by segment swapping for retrieval and discovery. In: CVPR (2022)
38. Shi, J., Malik, J.: Normalized cuts and image segmentation. TPAMI (2000)
39. Shi, J., Yan, Q., Xu, L., Jia, J.: Hierarchical image saliency detection on extended cssd. TPAMI (2015)
40. Shin, G., Albanie, S., Xie, W.: Unsupervised salient object detection with spectral cluster voting. In: CVPRW (2022)
41. Siméoni, O., Puy, G., Vo, H.V., Roburin, S., Gidaris, S., Bursuc, A., Pérez, P., Marlet, R., Ponce, J.: Localizing objects with self-supervised transformers and no labels. arXiv preprint arXiv:2109.14279 (2021)
42. Siméoni, O., Sekkat, C., Puy, G., Vobecky, A., Zablocki, É., Pérez, P.: Unsupervised object localization: Observing the background to discover objects. In: CVPR (2023)
43. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., Massa, F., Haziza, D., Wehrstedt, L., Wang, J., Darcet, T., Moutakanni, T., Sentana, L., Roberts, C., Vedaldi, A., Tolan, J., Brandt, J., Couprie, C., Mairal, J., Jégou, H., Labatut, P., Bojanowski, P.: DINOv3 (2025), https://arxiv.org/abs/2508.10104
44. Skurowski, P., Abdulameer, H., Błaszczyk, J., Depta, T., Kornacki, A., Kozieł, P.: Animal camouflage analysis: Chameleon database. Unpublished manuscript (2018)
45. Sun, Y., Wang, S., Chen, C., Xiang, T.Z.: Boundary-guided camouflaged object detection. arXiv preprint arXiv:2207.00794 (2022)
46. Voynov, A., Morozov, S., Babenko, A.: Object segmentation without labels with large-scale generative models. In: ICML (2021)
47. Wang, L., Lu, H., Wang, Y., Feng, M., Wang, D., Yin, B., Ruan, X.: Learning to detect salient objects with image-level supervision. In: CVPR (2017)
48. Wang, Y., Shen, X., Hu, S.X., Yuan, Y., Crowley, J.L., Vaufreydaz, D.: Self-supervised transformers for unsupervised object discovery using normalized cut. In: CVPR (2022)
49. Yan, Q., Xu, L., Shi, J., Jia, J.: Hierarchical saliency detection. In: CVPR (2013)
50. Yan, W., Chen, L., Kou, H., Zhang, S., Zhang, Y., Cao, L.: Ucod-dpl: Unsupervised camouflaged object detection via dynamic pseudo-label learning. In: CVPR (2025)
51. You, Z., Kong, L., Meng, L., Wu, Z.: FOCUS: Towards universal foreground segmentation. In: AAAI (2025)
52. Yuan, Y., Liu, W., Gao, P., Dai, Q., Qin, J.: Unified unsupervised salient object detection via knowledge transfer. arXiv preprint arXiv:2404.14759 (2024)
53. Zhang, Y., Wu, C.: Unsupervised camouflaged object segmentation as domain adaptation. In: ICCV (2023)
54. Zheng, P., Gao, D., Fan, D.P., Liu, L., Laaksonen, J., Ouyang, W., Sebe, N.: Bilateral reference for high-resolution dichotomous image segmentation. arXiv preprint arXiv:2401.03407 (2024)
55. Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., Kong, T.: ibot: Image bert pre-training with online tokenizer. arXiv preprint arXiv:2111.07832 (2021)
56. Zhu, W., Liang, S., Wei, Y., Sun, J.: Saliency optimization from robust background detection. In: CVPR (2014)