Weakly supervised salient object detection via double object proposals guidance

Zhiheng Zhou\textsuperscript{1,2} | Yongfan Guo\textsuperscript{1,2} | Ming Dai\textsuperscript{1,2} | Junchu Huang\textsuperscript{1,2} | Xiangwei Li\textsuperscript{3}

\textsuperscript{1} School of Electronic and Information Engineering, South China University of Technology, Guangzhou, China
\textsuperscript{2} Key Laboratory of Big Data and Intelligent Robot, Ministry of Education, South China University of Technology, Guangzhou, China
\textsuperscript{3} School of computer & Artificial Intelligence, Lanzhou Institute of Technology, Lanzhou, China

Correspondence
Zhiheng Zhou, School of Electronic and Information Engineering, South China University of Technology, Guangzhou, China.
Email: zhouzh@scut.edu.cn

Funding information
National Natural Science Foundation of China, Grant/Award Number: 61871188; National Key R&D Program of China, Grant/Award Number: 2018YFC0309400; Guangzhou city science and technology research projects, Grant/Award Number: 201802020008

Abstract
The weakly supervised methods for salient object detection are attractive, since they greatly release the burden of annotating time-consuming pixel-wise masks. However, the image-level annotations utilized by current weakly supervised salient object detection models are too weak to provide sufficient supervision for this dense prediction task. To this end, a weakly supervised salient object detection method is proposed via double object proposals guidance, which is generated under the supervision of double bounding boxes annotations. With the double object proposals, the authors’ method is capable of capturing both accurate but incomplete salient foreground and background information, which contributes to generating saliency maps with uniformly highlighted saliency regions and effectively suppressed background. In addition, an unsupervised salient object segmentation method is proposed, taking advantage of the non-parametric statistical active contour model (NSACM), for segmenting salient objects with complete and compact boundaries. Experiments on five benchmark datasets show that the authors’ weakly supervised salient object detection approach consistently outperforms other weakly supervised and unsupervised methods by a considerable margin, and even has comparable performance to the fully supervised ones.

1 | INTRODUCTION

As a dense prediction task, salient object detection is designed to identify the most prominent regions in various scenarios. Salient object detection thus has benefited a wide range of application fields, including image compression [1], object segmentation [2] and object recognition [3]. An enormous amount of salient object detection methods have been proposed, which can be roughly divided into unsupervised [4–6], weakly [7, 8] and fully supervised [9–11] approaches. Among these various methods, the weakly supervised approaches are attractive, since they release the burden of annotating time-consuming pixel-wise masks that are indispensable for the fully supervised ones, and show better performance than the unsupervised ones.

However, it is challenging to extract accurate and complete saliency regions in weak supervision settings. On the one hand, the image-level supervision sources (i.e. the category label and caption), utilized by current weakly supervised salient object detection methods [7, 8], are too weak to provide sufficient supervision for the inference of salient object detection. Furthermore, the image-level annotations lack the spatial information that is essential to the dense prediction tasks. As a result, these methods fail to uniformly highlight the whole salient object as shown in Figure 1(a). Compared with the fully supervised salient object detection models, there is still a large margin for improvement over these weakly supervised approaches. On the other hand, although an alternative weak supervision for weakly supervised dense prediction tasks is the bounding boxes annotation that contains both semantic and spatial information about objects, how to make full use of the bounding boxes supervision for weakly supervised salient object detection remains unexplored.

To solve the above problems, we propose a weakly supervised salient object detection method via double object proposals guidance. The bounding boxes annotations are utilized to supervise the generation of object proposals, which are a set
of candidate bounding boxes expected to cover all objects in the image and can be converted to be effective guidance cues as proved in previous proposal-based salient object detection methods [4–6]. For the first time, we exploit double object proposals guidance for accurate and complete salient object detection based on two main considerations.

Firstly, object proposals are supposed to provide both accurate but incomplete salient foreground and background information. Nevertheless, we observe that the outer object proposals (OPs), that is the conventional object proposals that tightly touch the boundary of the objects, can only provide vague salient foreground information. As shown in Figure 1(c), the region outside the OPs is the accurate but incomplete background, while that inside the OPs is composed of both foreground and background, which provides ambiguous guidance for salient object detection. Therefore, we additionally exploit the inner object proposals (IPs), that is the proposals expected to be inside the objects, to capture accurate but incomplete foreground information as shown in Figure 1(d).

Secondly, the guidance information converted by object proposals should be as confident as possible. To this end, instead of directly using the conventional object proposal generators [12, 13], we adopt currently supervised object detection models [14] benefited from the powerful convolutional neural networks (CNNs) for high-quality double object proposals (IPs and OPs). These object proposals are generated under the supervision of the double bounding boxes annotations (i.e. the inner and outer bounding boxes as shown in Figure 1(b)).

The confident double object proposals are first processed to provide guidance for weakly supervised salient object detection, which enables our method to capture both accurate but incomplete salient foreground and background information. Then, we jointly consider the high-level guidance information, mid-level cues of superpixels and low-level pixel-wise colour, thereby yielding reliable probabilities of salient foreground and background by the Gaussian mixture model (GMM). After that, we design a simple but efficient “soft” saliency assignment to assign saliency values for each pixel, using the difference between salient foreground and background probabilities. At last, our weakly supervised salient object detection method is able to uniformly highlight saliency region meanwhile effectively suppress the cluttered background.

The main contributions of our work are summarized as:

- We are the first to exploit the double object proposals for weakly supervised salient object detection. The double object proposals are capable of providing accurate but incomplete salient foreground and background information, which enables our method to achieve comparable performance to the fully supervised salient object detection models.
- We jointly consider the high-level guidance information, mid-level cues of superpixels and low-level pixel-wise colour by the Gaussian mixture model, as well as design a soft saliency assignment to assign saliency values for each pixel. Therefore, our method is able to uniformly highlight saliency regions meanwhile effectively suppress the cluttered background.
- We additionally expand this paper to salient object segmentation and propose an unsupervised salient object segmentation method (named as AcSeg). Taking advantage of the non-parametric statistical active contour model (NSACM), which can handle objects with complex shapes and topological changes, our AcSeg can segment salient objects with much more compact and complete boundaries.

The rest of the paper is organized as follows: Section 2 reviews some representative related works. Sections 3 and 4, respectively, describe our proposed weakly supervised salient object detection approach and unsupervised salient object segmentation method in detail. In Section 5, the training of the object detection model and some parameters of our method are provided. Section 6 shows the experimental comparisons and some analysis. Section 7 draws the conclusion.

## 2 RELATED WORK

### 2.1 Salient object detection

The traditional unsupervised salient object detection methods focus on some heuristic priors, for example boundary prior [15, 16], contrast prior [17] and background connectivity prior [18]. In [15], Li et al. extract the image boundaries as background templates. Based on the templates, dense and sparse appearance models are constructed to compute the pixel-level saliency. Cheng et al. [17] compute the global contrast differences and spatial relevancies to determine the regional saliency. In [18], Zhu et al. propose an effective boundary connectivity measure, which characterizes the spatial layout of different superpixels with respect to the image boundary and can be integrated with other prior cues for salient object detection. Although these heuristic saliency cues promote the performance of salient object detection methods to some extent, they are too fragile to handle images with much more complicated scenes.

Benefited from the powerful convolutional neural networks (CNNs), which are capable of extracting different features for different tasks, a large number of salient object detection models [9–11] based on deep learning have been proposed. In [9], Hou et al. introduce the short connections to the skip-layer structures for the purpose of providing more advanced
representation at the layer of CNNs. Zhang et al. [10] aggregate multi-level convolutional features in a flexible manner to provide accurate salient object labelling for their saliency model. Zhang et al. [11] aim at learning uncertain convolutional features for accurate salient object detection. Though these supervised models have shown superior performance, they severely rely on costly pixel-wise annotations.

To release the burden of annotating time-consuming pixel-wise masks, weakly supervised learning has been introduced in dense prediction tasks. In weakly supervised salient object detection, some attempts [7, 8] leveraging the image-level weak supervisions have been made. Wang et al. [7] utilize the category labels to develop a weakly supervised learning method for salient object detection. A foreground inference network and a global smooth pooling layer are designed for weakly supervised learning. In [8], Zeng et al. propose a unified framework to train saliency models with diverse weak supervision sources, that is the category labels and captions. Nevertheless, these methods fail to uniformly highlight the whole salient objects due to the weakness of the image-level annotations.

2.2 Object proposals

The pioneering work of Alexe et al. [12] proposes a generic objectness measure to generate object proposals containing objects of any class. The EdgeBoxes [13] generates class-independent object proposals by directly using edges. To incorporate high-level information, these proposals are utilized to facilitate salient object detection. Early salient object detection methods [4, 5] generate the objectness maps by simply merging thousands of proposals based on their objectness scores. The objectness maps can roughly indicate the location of potential saliency regions, and serve as effective prior cues for salient object detection. Different from these methods, Huang et al. [6] utilize the EdgeBoxes to generate a series of object proposals, then take the proposals as the bags of instances of multiple instances learning. However, these methods are not only limited by the performance of the object proposal generators, but they can only provide ambiguous salient foreground information for salient object detection.

2.3 Salient object segmentation

Salient object segmentation, typically followed after the salient object detection, aims at generating high-quality binary masks, where the ones should be assigned to pixels belonging to salient objects, as well as the zeros to the rest of the pixels. In FT [19], Achanta et al. first utilize the mean-shift segmentation algorithm to over-segment the image into subregions with well-preserved boundaries. Then they retain the subregions whose average saliency is greater than an adaptive threshold. The SalCut [17] method extends the GrabCut [20] algorithm and automatically initializes it using saliency maps, to iteratively refine the segmentation results. In [21], Jiang et al. exploit the object-level shape prior and incorporate it into an iterative energy minimization framework, resulting in binary segmentation of the salient objects. As evaluated in [17], the SalCut method shows superior performance and perform efficiently among these salient object segmentation approaches. However, considering less the boundary information, it may yield incomplete salient objects and cannot handle images containing widely separated multiple salient objects.

3 Weakly supervised salient object detection

An overview of our weakly supervised salient object detection method is shown in Figure 2. First of all, we train the object detection method R-FCN [14] under the supervision of the double bounding boxes labels, which enables the R-FCN to generate our desired double object proposals (IPs and OPs). Based on the double object proposals, the regions inside and outside the salient object (RIS and ROS) are extracted, which are considered as the source of accurate but incomplete guidance information about salient foreground and background. Then we expand these regions by exploiting both the colour and spatial correlations between different superpixels to extract the initial foreground and background regions (IFR and IBR). After that, two different Gaussian mixture models (GMMs) are built for salient foreground and background respectively, to integrate multi-level information. Finally, a soft saliency assignment is designed to assign saliency for each pixel.

3.1 Extraction of regions inside and outside salient object

Given an input image, the well-trained object detection model R-FCN generates a few hundred proposals with different objectness scores, which may contain our desired double object proposals, that is inner object proposals (IPs) and outer object proposals (OPs). However, since the R-FCN model is designed for the generation of conventional bounding boxes with specific categories, directly using the objectness scores of proposals to extract regions inside and outside the salient object is not confident enough. Especially, the objectness scores of IPs can greatly mislead the extraction of the region inside the salient object as shown in Figure 3(b).

To solve this problem, we weight an IP’s score by using the maximum among the objectness scores of OPs completely containing the IP. The reasons are two-fold. First, we observe that the objectness scores of OPs are more reliable than those of the weird IPs, which will be validated in the experiments (as shown in Figure 4). Second, since we expect the IPs and OPs to be inside and outside the salient object respectively, the IPs should be fully contained by corresponding OPs.

In detail, we first select \(N_I\) outer object proposals and \(N_O\) inner object proposals with the highest objectness scores. Then we normalize these objectness scores to the range of [0, 1]. And
**FIGURE 2** The overall framework of our weakly supervised salient object detection method. From left to right (columns): input image, double bounding boxes supervision for object detection model, input image with inner object proposals (IPs) and outer object proposals (OPs) overlaid, the regions inside and outside the salient object (RI and ROS), the initial foreground and background regions (IFR and IBR) and the final saliency map.

**FIGURE 3** The illustration of different score maps. (a) is the input images with double object proposals overlaid. And the score maps generated by non-weighted scores of IPs, weighted scores of IPs and scores of OPs are shown in (b), (c) and (d) respectively. Comparing (b) and (c), it is obvious that the weighted operation in Equation (1) can remedy the score maps assigned by the badly detected IPs.

The weighted scores of IPs are defined as

$$w_{SI} = s_i \cdot \exp(2 \cdot s_{O_{\text{max}}}),$$  

where $s_i$ is the normalized score of $IP_i$, and $s_{O_{\text{max}}}$ is the maximum among the normalized score of OPs completely containing current $IP_i$. After that, we re-normalize the weighted scores of IPs, and the score maps of IPs and OPs are thus obtained

$$\text{SiMap}(x) = \min \left\{ 1, \sum_{i=1}^{N_I} w_{SI} \cdot \delta_i(x) \right\},$$

$$\text{SoMap}(x) = \min \left\{ 1, \sum_{i=1}^{N_O} s_{Oi} \cdot \delta_i(x) \right\}.$$  

**FIGURE 4** The illustration of initial foreground and background regions extraction. From left to right: input image with superpixel boundaries overlaid, input image with RIS and ROS overlaid, the foreground connectivity map (FGC), the initial foreground region (IFR), the background connectivity map (BGC) and the initial background region (IBR).
where \( \delta_i(x) \) is 1 if corresponding object proposal (IP_i or OP_i) includes pixel \( x \).

The score maps generated by non-weighted and weighted scores of IPs are illustrated in Figures 3(b) and 3(c), respectively, proving that the badly assigned score maps of IPs can be remedied by our weighted operation in Equation (1).

With the more reliable score maps of IPs (SiMap), the region inside the salient object (RIS) is obtained by simply thresholding the SiMap with a fixed high threshold of \( \Theta^{IP}_h \). While the region outside the salient object (ROS) is determined by an adaptive threshold. To be specific, we threshold the score map of OPs (SiMap) with thresholds from large to small, until the binary SoMap can completely include the binary SiMap with a fixed low threshold of \( \Theta^{IP}_l \). Then the complement of the binary SoMap is considered as the region outside the salient object (ROS). Note the adaptive thresholding of SoMap can exclude as much background as possible while guaranteeing full inclusion of foreground by the binary SoMap.

### 3.2 Initial foreground and background regions extraction

As shown in the second image of Figure 5, the regions inside and outside the salient object (RIS and ROS) contain salient foreground and background information respectively. Taking the colour similarity together with the spatial location into consideration, we adopt the “connectivity” of different superpixels to the RIS and ROS to expand these regions and thus extract initial foreground and background regions.

At first, we utilize the SLIC algorithm [22] to oversegment the input image \( \Omega \) into a set of superpixels \( S = \{ s_1, s_2, \ldots, s_k \} \). Then the mean normalized colour in LAB colour space and the normalized coordinates of pixels are utilized to represent the features of each superpixel by \( F_s(s_i) = \{ L(s_i), A(s_i), B(s_i), X(s_i), Y(s_i) \} \). Thus an undirected weighted graph based on these features is constructed, and the weight between adjacent superpixels \( (s_i, s_j) \) is defined as

\[
    w(s_i, s_j) = \left\| F_s(s_i) - F_s(s_j) \right\|^2.
\]

According to [18], the foreground and background connectivity maps are defined as

\[
    \begin{align*}
    FGC(s_i) &= \frac{\sum_{j=1}^{k} C(s_i, s_j) \cdot \delta(s_j \in RIS)}{\sqrt{\sum_{j=1}^{k} C(s_i, s_j)}}, \\
    BGC(s_i) &= \frac{\sum_{j=1}^{k} C(s_i, s_j) \cdot \delta(s_j \in ROS)}{\sqrt{\sum_{j=1}^{k} C(s_i, s_j)}},
    \end{align*}
\]

where \( \delta(\cdot) \) is 1 if superpixel is totally included by corresponding regions (RIS or ROS), and the operand \( C(s_i, s_j) \) is calculated as

\[
    C(s_i, s_j) = \exp\left(-\frac{g^2(s_i, s_j)}{2\sigma_{clr}^2}\right),
\]

where the geodesic distance \( g(s_i, s_j) \) between superpixels \( s_i \) and \( s_j \) is defined as the accumulated edge weights along their shortest path on the graph, and \( \sigma_{clr} \) is set to 10 according to [18]. Then the foreground and background connectivity maps (FGC and BGC) are normalized to the range of \([0, 1]\).

Finally, the initial foreground and background region (IFR and IBR) are obtained by extracting the regions with maximum value in corresponding connectivity maps as shown in Figure 5. Note in Equation (3), we use both LAB colour and spatial coordinates to calculate the weight of the constructed graph, which can guarantee the IFR and IBR to be the accurate but incomplete foreground and background to some degree.

### 3.3 Saliency assignment

Inspired by the well-known GrabCut [20] method, which assigns two different Gaussian mixture models (GMMs) for foreground and background, we utilize the initial foreground and background regions (IFR and IBR) to construct these two kinds of GMMs. The number of components of each GMM is typically set to \( N = 5 \). For each component \( n \), we need to estimate the mixture weighting coefficient \( \omega_n \), mean value \( \mu_n \) and covariance matrix \( \Sigma_n \).
FIGURE 6 The illustration of qualitative comparisons between our AcSeg and other salient object segmentation methods. From left to right (columns): our input saliency maps with scaling factor $r = 55$, the initial contour of AcSeg, our AcSeg, the baseline (BS) (i.e. the saliency maps with $r = 35$ binarized by the Otsu method [31]), FT [19], SalCut [17] and ground truth.

In order to jointly consider multi-level information, we utilize the RGB colour of pixel $x$ and the mean LAB colour of superpixels to represent the features for GMMs by $F_S(x) = \{R(x), G(x), B(x), L(x_s), A(x_s), B(x_s)\}$, where $x_s$ represents the superpixel containing pixel $x$. Such six-dimensional features and the initial regions (IFR and IBR) can force the GMMs to simultaneously integrate the high-level guidance information, mid-level colour of superpixels and low-level pixel-wise colour. Then we adopt the K-means algorithm to respectively cluster the pixels in IFR and IBR into $N$ clusters based on the six-dimensional features $F_S$. For a given GMM component $n$ in IFR (or IBR), the pixel set $P_{n}$ of this component is determined. Thus the mean value $\mu_n$ and covariance $\Sigma_n$ are estimated as the sample mean and covariance of feature values in $P_{n}$, and the mixture weighting coefficient $\omega_n$ is calculated as,

$$\omega_n = \frac{|P_{n}|}{\sum_{n=1}^{N} |P_{n}|}.$$  \hspace{1cm} (6)

We adopt the “data term” in GrabCut [20] to calculate the salient foreground probability of pixel $x$ via using the estimated parameters from IFR, which is defined as

$$P^F(x) = \frac{1}{1 + \exp \left( r \cdot D(x) \right)}.$$  \hspace{1cm} (8)

As shown in Figure 6, the soft saliency assignment with different scaling factors in Equation (8) results in different saliency maps. The impact of different scaling factors on salient object detection performance will be further analyzed in the experimental part. Unless otherwise specified, the salient object detection results in the experiments are based on $r = 55$. And the background probability ($P^B$) is similarly obtained from the parameters estimated from the IBR.

With the reliable probabilities of salient foreground and background ($P^F$ and $P^B$), there are various alternative methods to distinguish salient objects from the background. Rather than using the less efficient iterative optimization algorithm [23], which can generate hard segmentation results [20], we design a “soft” saliency assignment. The difference between salient foreground and background probabilities is computed by $D(x) = P^F(x) - P^B(x)$. It is obvious that the larger the $D(x)$, the larger saliency value should be assigned to the pixel $x$. Although a simple and straightforward idea is to normalize the difference value directly to $[0, 1]$, this kind of linear scaling will blur the difference between foreground and background as shown in the second image of Figure 6. Therefore, the difference is firstly normalized to the range of $[-1, 1]$, then a non-linear scaling via the modified sigmoid function with a scaling factor $r$ is utilized as

$$S(x) = \frac{1}{1 + \exp \left( r \cdot D(x) \right)}.$$  \hspace{1cm} (8)
4 | UNSUPERVISED SALIENT OBJECT SEGMENTATION

In this section, we will detail our unsupervised salient object segmentation method, that is non-parametric statistical active contour model (NSACM)-based salient object segmentation method (AcSeg), which is capable of extracting salient objects with complete and compact boundaries.

The active contour model (ACM) owns the unique property of handling objects with complex shapes and topological changes [24], which has the potential to yield binary saliency maps with fine boundaries. As an important branch of the ACM, the NSACM maintains the unique property of the ACM and is easier to be embedded prior distribution information. Recently, a prior distribution active contour model (PACM) [25] based on NSACM is proposed, which is a framework that can flexibly integrate prior distribution information to enhance the segmentation ability of the NSACM models. Previous NSACM-based methods only rely on intensity values of the image itself to construct different energy functions to measure the distribution discrepancy within different regions. In PACM, the prior distributions of foreground and background are introduced and a prior energy term is thus constituted. Therefore, with the level set function [29] \( \phi \), the total energy function is defined as

\[
E(p, q, \beta', q', \phi) = E_{\text{AMP}}(p, q, \phi) + E_{\text{Prior}}(p, q, \beta', q', \phi),
\]

where \( p \) and \( q \) are the estimated distributions of the segmented foreground and background, while the \( \beta' \) and \( q' \) are corresponding prior distributions. These distributions are based on the intensity value \( i \) in the intensity domain. Given the segmented foreground and background regions (FR and BR), the distributions \( p \) and \( q \) are estimated via the Parzen density estimation [30] as

\[
p(i) = \frac{1}{|\text{FR}|} \int_{\text{FR}} K_\sigma(i - I(x)) dx,
\]

\[
q(i) = \frac{1}{|\text{BR}|} \int_{\text{BR}} K_\sigma(i - I(x)) dx,
\]

where \( K_\sigma \) is the Gaussian kernel function with bandwidth \( \sigma \).

In Equation (9), the \( E_{\text{AMP}}(p, q, \phi) \) is the average misclassification probability (AMP) energy [26], and the \( E_{\text{Prior}}(p, q, \beta', q', \phi) \) is the prior energy term based on the Bhattacharyya distance [28]. The specific expressions of the AMP and the prior terms as well as their corresponding gradient descent flows (i.e. \( \frac{dE}{d\phi} \) AMP and \( \frac{dE}{d\phi} \) Prior) can be found in [25].

Here, we improve the prior distribution active contour model (PACM) [25] in three aspects, making it achievable to segment salient objects more accurately. These improvements are shown below.

Firstly, the prior distributions are automatically determined by the saliency maps. On the contrary, these distributions need to be manually selected in PACM. In a well-assigned saliency map, the high saliency values point to saliency regions, while the low ones indicate background. Therefore, in a normalized saliency map \( S \) with the range of \([0, 1]\), the region with saliency values greater than a high threshold \( \theta_h \) (HR) can be considered as the source of the prior foreground distribution \( \beta' \). Similarly, the region with saliency values lower than a low threshold \( \theta_l \) (LR) can be viewed as the source of the prior background distribution \( q' \). The prior distributions \( \beta' \) and \( q' \) are thus obtained as

\[
\beta'(i) = \frac{1}{|\text{HR}|} \int_{\text{HR}} K_\sigma(i - I(x)) dx,
\]

\[
q'(i) = \frac{1}{|\text{LR}|} \int_{\text{LR}} K_\sigma(i - I(x)) dx.
\]

Secondly, a new initialization method is adopted. As pointed out in [26], subject to the independent and identical distribution (i.i.d) assumption, which is not always true in real images, the optimization of designed energy functions does not guarantee a global optimal solution. The initial contour is essential for these i.i.d-based non-parametric statistical active contour models, because we can only get a local solution most of the time. To this end, previous NSACM-based methods need to manually determine a rectangular frame or ellipse close to the object to be segmented for more accurate real image segmentation. We utilize the binary saliency map binarized by the Otsu [31] algorithm to automatically initialize the level set function \( \phi \). This initialization ensures that the initial contour is near the boundary of the saliency object as shown in the second column in Figure 7. Therefore, even the optimization is trapped in a local solution in the correct direction, we can also obtain more satisfactory segmentation results.

Thirdly, the saliency information is added to the total gradient descent flow. For the computation efficiency, the non-parametric statistical active contour models only use the intensity information to estimate the distributions. However, this loss of the original colour information can confuse the difference between the foreground and background in some scenarios, since only 256 bins of the histogram are exploited [32]. Here, to improve the stability of our AcSeg

---

**Algorithm 1 AcSeg algorithm**

Input: Saliency map \( S \) and intensity image \( I \).

- Set the maximum iteration number \( T_{\text{max}} \);
- Compute the prior foreground and background distributions \((\beta' \text{ and } q')\) according to Equation (11);
- Obtain the initial level set \( \phi_0 \) by adaptively thresholding the saliency map \( S \);

**while** \( t < T_{\text{max}} \) **do**

- Estimate the distributions \((\beta \text{ and } q)\) inside and outside the contour determined by \( \phi \) according to Equation (19);
- Compute the total gradient flow \( \frac{d\phi}{dt} \) according to Equation (14);
- Update level set \( \phi \rightarrow \phi_{t+1} \) according to Equation (16);
- **End**

---
and keep it efficient, we add another gradient descent flow by converting the saliency map $S$ to be signed values as
\[
\frac{\partial E}{\partial \phi} = -\delta(\phi(x))[0.5 - S(x)],
\]
where the approximated delta function $\delta$ is defined as
\[
\delta(z) = \frac{1}{\pi} \frac{\epsilon}{\epsilon^2 + z^2}.
\]

Therefore, the total gradient descent flow is obtained
\[
\frac{\partial E}{\partial \phi} = \left[ \frac{\partial E}{\partial \phi} \right]_{\text{AMP}} + \left[ \frac{\partial E}{\partial \phi} \right]_{\text{Prior}} + \left[ \frac{\partial E}{\partial \phi} \right]_{\text{Sal}} + \left[ \frac{\partial E}{\partial \phi} \right]_{\text{Length}},
\]
where the last term is the gradient flow of the curve length penalty term [33], which is defined as
\[
\frac{\partial E}{\partial \phi} = -\lambda \delta(\phi(x)) \text{div} \left( \nabla \phi(x) \right),
\]
Finally, the update of the level set $\phi$ at step $t$ is computed as
\[
\phi_{t+1}(x) = \phi_t(x) - \Delta_t \cdot \frac{\partial E}{\partial \phi},
\]
where $\Delta_t$ is timestep controlling the evolution speed of level set $\phi$. The overall process of our AcSeg is summarized in Algorithm 1.

5 | IMPLEMENTATION DETAILS

To train the supervised object detection model R-FCN [14], double bounding boxes labels (i.e. inner bounding boxes and outer bounding boxes) supervising the generation of inner object proposals (IPs) and outer object proposals (OPs) are needed. However, because we need two kinds of category-agnostic bounding boxes, current public object detection datasets do not meet our requirement. Therefore, the large-scale salient object detection dataset DUTS [7], which contains 10,553 training images and 5019 test images, is utilized to generate double bounding boxes labels. More specifically, as shown in Figure 8, we treat the minimum circumscribed rectangles of salient objects in the binary masks as the outer bounding boxes, and the maximum inscribed rectangles of salient objects are considered as the inner bounding boxes.

We directly utilize the well-trained R-FCN model1 from the Tensorflow Detection Model Zoo to initialize our model. This R-FCN model is based on ResNet101 [34] and trained on the COCO [35] dataset. Then, we utilize the Adam optimizer with an initial learning rate of 1e-7 to fine-tune the R-FCN model. We additionally train the R-FCN model for 60 epochs with an NVIDIA Titan X GPU. The learning rate is divided by 10 every 20 epochs. The number of classes is set to 2 and the batch size is 1. Other parameters are the same as the config file2 in this Zoo.

There are some important parameters, including the number of the object proposals (i.e. $N_i$ and $N_o$), some thresholds (i.e. $\theta_{IP}^h$, $\theta_{IP}^l$, $\theta_{Ac}^h$ and $\theta_{Ac}^l$) and the scaling factor $r$ in Equation (8). To determine these hyperparameters, we conduct several sets of experiments in Section 6.4.3.

In the unsupervised salient object segmentation part (Section 4), a set of parameters, that is the $\varepsilon$ in approximated delta function (Equation 13), the $\lambda$ in the curve length penalty term (Equation 15), the timestep $\Delta_t$ (Equation 16) and the maximum iteration number $T_{\text{max}}$ are not determined yet. Referring to previous NSACM methods [25–28], we empirically obtain a set of fixed parameters. More specifically, in order to converge as quickly as possible, we set a relatively larger timestep (i.e. $\Delta_t = 5$), so a relatively small maximum iteration number can be set (i.e. $T_{\text{max}} = 15$). According to [33], we set $\varepsilon = 1$ and $\lambda = 0.001 \times 255 \times 255$. The experimental results demonstrate that this set of parameters makes our AcSeg have comparable performance to the SalCut [17] method.

6 | EXPERIMENTS

6.1 | Datasets and evaluation metrics

We evaluate the performance of our approach on five public benchmark datasets, including SED [36], DUTOMRON [16], PASCAL-S [37], ECSSD [38] and THUR [39]. The SED dataset is composed of two subsets: SED1 and SED2. The SED1 contains 100 images with only one salient object, while two salient objects are in each image of the SED2. The DUTOMRON dataset has 5186 images containing one or more salient objects. The PASCAL-S dataset contains 850 images with cluttered scenes and is one of the most challenging salient object
detection datasets. The ECSSD dataset has 1000 images with complex structures. The THUR is another large-scale dataset containing 6232 pixel-wise labelled annotations.

Mean absolute error (MAE), F-measure ($F_\beta$), and precision-recall (PR) curves are utilized to evaluate the performance of different salient object detection models. The PR curve is calculated by setting different thresholds ranging from 0 to 1 to binarize the saliency maps and comparing these binary saliency maps with the ground truths. The $F_\beta$ is a comprehensive measure of precision and recall, which is defined as

$$F_\beta = \frac{(1 + \beta^2) \cdot \text{Precision} \cdot \text{Recall}}{\beta^2 \cdot \text{Precision} \cdot \text{Recall}},$$

where $\beta^2 = 0.3$ to emphasize precision over recall. In the comparison experiments, this paper examines the maximum $F_\beta$ ($F_m$), which is computed from the PR curves and is a good summary of the salient object detection performance [42]. The MAE is computed as the average pixel-wise absolute discrepancy between the saliency map $S$ and the ground truth $G$,

$$\text{MAE} = \frac{1}{W \times H} \sum_i \sum_j |S(i,j) - G(i,j)|,$$

where $W$ and $H$ are the width and height of the input image.

### 6.2 Performance comparisons in salient object detection

We compare our weakly supervised salient object detection model with eight state-of-the-art methods, including three fully supervised methods (Amulet [10], UCF [11] and DRFI [40]), four unsupervised methods (MR [16], SO [18], MBS [41] and DSR [15]), and one weakly supervised method (MWS [8]). For fair comparisons, the saliency maps of these methods are provided directly by the authors or generated by their released codes.

The quantitative comparisons between our proposed approach and eight state-of-the-art salient object detection models on five datasets are shown in Table 1. It can be observed that our method consistently outperforms other weakly supervised and unsupervised methods by a considerable margin. Furthermore, our method even has comparable performance to the CNN-based supervised models (i.e. Amulet [10] and UCF [11]). It is worth mentioning that our method shows superior performance in the challenging datasets, that is DUTOMRON, PASCAL-S and THUR. In addition, we draw PR curves of different methods on three popular datasets, that is DUTOMRON, PASCAL-S and ECSSD. As shown in Figure 9, though our method performs worse than the CNN-based approaches when the recall is much higher, it can easily reach higher precision when the recall is relatively low. This is because we additionally exploit the inner object proposals (IPs) for capturing accurate but incomplete salient foreground information, while the precision of IPs is high, which can be reflected from the precision of RIS as shown in Figure 4.

A few qualitative comparisons are illustrated in Figure 10. Note, we choose these images with a variety of different circumstances, including salient objects with different sizes, different contrast between background and foreground, different number of salient objects. From the comparison results, we can observe that our method is robust enough to handle images with different scenes.
TABLE 1 \( F_m \) (higher is better) and MAE (lower is better) comparisons between our proposed salient object detection method and other eight state-of-the-art methods. Note the Amulet [10], UCF [11] are CNN-based supervised methods. The best three results are shown in red, green, and blue fonts.

|       | THUR | PASCAL-S | ECSSD | DUTOMRON | SED1 | SED2 |
|-------|------|----------|-------|----------|------|------|
| Amulet | 0.773 | 0.094    | 0.841 | 0.100    | 0.927 | 0.059 |
| UCF    | 0.753 | 0.112    | 0.829 | 0.115    | 0.920 | 0.069 |
| DRFI   | 0.741 | 0.150    | 0.741 | 0.226    | 0.843 | 0.164 |
| MR     | 0.685 | 0.179    | 0.727 | 0.221    | 0.801 | 0.190 |
| SO     | 0.674 | 0.150    | 0.716 | 0.199    | 0.783 | 0.171 |
| MBS    | 0.683 | 0.162    | 0.721 | 0.196    | 0.789 | 0.171 |
| DSR    | 0.661 | 0.142    | 0.688 | 0.204    | 0.776 | 0.171 |
| MWS    | 0.767 | 0.098    | 0.819 | 0.133    | 0.904 | 0.096 |
| Ours   | 0.792 | 0.084    | 0.847 | 0.101    | 0.905 | 0.078 |

FIGURE 9 The precision-recall (PR) curves on three popular salient object datasets. Note our method can easily reach high precision when the recall is relatively low.

6.3 Performance evaluation in salient object segmentation

In this section, we compare our AcSeg with other salient object segmentation methods, that is FT [19] and SalCut [17]. The saliency maps generated by our proposed weakly supervised salient object detection approach with scaling factor \( r = 55 \) are feed to these segmentation methods. Then, the \( F_\beta \) and MAE of different segmentation results are computed. In addition, since the saliency maps assigned by our soft saliency assignment with larger scaling factor \( r \) are nearly binarized, the saliency maps with \( r = 355 \) binalized by the Otsu method [31] is viewed as the baseline (BS).

The quantitative comparisons are shown in Figure 11. It is obvious that the BS performs better than the FT method, verifying the high-quality of saliency maps generated by a larger scaling factor. The experiment results also show our AcSeg and SalCut outperform the BS by a large margin, and our AcSeg has comparable performance with the SalCut. The unique advantage of AcSeg is shown in Figure 7. Benefited from the NSACM, our AcSeg can generate salient objects with much more complete and compact boundaries. On the contrary, the SalCut...
The quantitative comparisons between our AcSeg and other salient object segmentation methods, along with the baseline (BS) yields incomplete salient objects and cannot handle images containing widely separated multiple salient objects. However, since AcSeg extremely relies on the quality of the saliency map, some wrongly highlighted parts will greatly mislead the segmentation. Therefore, some scattered background regions are retained in the segmentation results. Nevertheless, we believe that the performance of AcSeg can be further improved by enforcing consistency of saliency regions.

6.4 Analysis

6.4.1 Effectiveness of extracting RIS and ROS

We conduct experiments on different datasets to evaluate the effectiveness of extracting regions inside and outside the salient object (RIS and ROS), which are essential to provide accurate but incomplete guidance information about salient foreground and background. As shown in Figure 4, we compute the precision of RIS and the recall of the complement of ROS (i.e. the region expected to totally contain salient objects). The histograms show that the recall of ROS is much higher than the precision of RIS, indicating the rationality of weighting the objectness scores of IPs by those of corresponding OPs. In addition, the precision of RIS that directly generated by non-weighted objectness scores is also shown in Figure 4, which proves that the weighted operation in Equation (1) can greatly improve the accuracy of detecting region inside the salient object.

6.4.2 Ablation analysis of salient object detection

We perform two ablation experiments in Table 2. In the first experiment, we directly feed the regions inside and outside the salient object (RIS and ROS) generated by the score maps of IPs and OPs to GMMs. This combination causes a sharp drop in performance (the first row in Table 2), which means that our initial foreground and background regions (IFR and IBR) retain the accurate but incomplete property of RIS and ROS, meanwhile indeed provide more guidance information about salient foreground and background. In the second experiment, we confirm the insight that multi-level information is indispensable for the performance improvement of salient object detection. Without the mid-level information of superpixels, the performance of salient object detection is reduced (the second row in Table 2).

6.4.3 Parameters analysis

To verify the effect of the soft saliency assignment with different scaling factors \( r \) on the performance of salient object detection, we set four values \( r = 5, 55, 155 \) and \( 355 \) to obtain corresponding maximum \( F_\beta (F_m) \) and adaptive \( F_\alpha (F_a) \). The \( F_a \) is obtained by an image-dependent adaptive threshold \[19\] computed as twice as the mean saliency value of the saliency map and can reflect the spatial consistency of the saliency map. As shown in Figure 12, there exists an obvious trend, that is a larger \( r \) increases the \( F_a \) of the assigned saliency map, meanwhile decreases the \( F_m \) of that. This trend indicates that the soft saliency assignment with a larger \( r \) contributes to better uniformly highlighting saliency regions. However, this kind of saliency assignment can inevitably reduce the reliability of the assigned saliency maps. Hence, considering the trade-off between saliency maps with uniformly highlighted and reliable saliency regions, we choose \( r = 55 \) for our salient object detection.

The setting of the number of object proposals (i.e. \( N_{i} \) and \( N_{o} \)) should cover almost all salient objects while reducing...
redundancy as much as possible. To this end, we first count the number of separated salient objects in different images in the DUTS [7] dataset. Then, the proportion of images with different numbers of salient objects to the entire dataset is obtained as shown in Figure 13. It can be observed that almost all images contain no more than 3 salient objects. Recalling that we first normalize the objectness score of the selected object proposals, the lowest objectness score is thus normalized to 0. In this case, the actual number of object proposals is one less than the set number. Based on the above discussion, we set $N_I = N_O = 4$.

To determine different thresholds (i.e. $\theta_{IP}^h$, $\theta_{IP}^l$, $\theta_{Ac}^h$ and $\theta_{Ac}^l$), we conduct a series of experiments on the PASCAL-S dataset. Note the thresholds $\theta_{IP}^h$ and $\theta_{IP}^l$ are utilized to generate the regions inside and outside the salient object (RIS and ROS). The $\theta_{IP}^h$ should be slightly lower than $\theta_{IP}^l$, so that the ROS is pure background composition. Therefore, we can set $\theta_{IP}^l = \theta_{IP}^h - 0.1$. As shown in Table 3, we calculate the recall of ROS and the precision of RIS under different $\theta_{IP}^h$, which shows that different thresholds have little effect on these results. We set $\theta_{IP}^h = 0.95$ to ensure high reliabilities of RIS and ROS. As for the $\theta_{Ac}^h$ and $\theta_{Ac}^l$, which are used to determine the prior distributions of foreground and background, we can set $\theta_{Ac}^l = 1 - \theta_{Ac}^h$ since $\theta_{Ac}^l$ should be lower while $\theta_{Ac}^h$ should be higher. The impact of these thresholds on the performance of AcSeg in terms of $F_\beta$ and MAE is shown in Table 4. It can be concluded that AcSeg is not sensitive to these parameters. We set $\theta_{Ac}^h = 0.9$ for the reliability consideration.

6.4.4 Limitation analysis

We exploit the double object proposals to capture accurate but incomplete salient foreground and background information for weakly supervised salient object detection. However, we find that the insufficient guidance information about salient foreground and background can result in unsatisfactory saliency maps, which may be caused by two cases. First, inadequate salient foreground information is obtained when the salient object is extremely irregular (e.g. the hollow bicycle in the first row of Figure 14). In this case, our method can neither highlight the foreground uniformly nor suppress the background effectively. Second, when the salient object is nearly full of the whole image as shown in the second row of Figure 14, no or little background information can be obtained, which makes our method incompetent to remove enough background.

![Figure 14](image-url) The failure examples of our salient object detection. (a) Input images; (b) Failure salient object detection results; (c) Ground truths.
6.5 | Computation cost

We implement the AcSeg and R-FCN using Python with Tensorflow [43], on our computer with a 3.0 GHz CPU, 16 GB RAM and an NVIDIA Titan X GPU. On the ECSSD dataset, the AcSeg and R-FCN respectively take 0.25 and 0.08 s to process each image on average. Other parts of our salient object detection are only implemented using Python on the CPU. They cost another 0.29 s.

7 | CONCLUSION

This paper proposes a weakly supervised salient object detection method via double object proposals guidance. The double object proposals are generated under the supervision of the double bounding boxes, and are able to provide accurate but incomplete salient foreground and background information for salient object detection. Considering jointly high-level guidance information, mid-level features and low-level pixel-wise colour, our method can yield saliency maps with uniformly highlighted saliency region meanwhile effectively suppressed background. In addition, we expand this paper to salient object segmentation and propose an unsupervised salient object segmentation method (AcSeg). Taking advantage of NSACM that can handle objects with complex shapes and topological changes, our AcSeg is capable of segmenting salient objects with complete and compact boundaries. Experiments on five benchmark datasets show that our weakly supervised salient object detection approach consistently outperforms other weakly supervised and unsupervised methods by a considerable margin, and even has comparable performance to the fully supervised ones.

ACKNOWLEDGEMENTS

The work is supported by National Key R&D Program of China (2018YFC0309400), National Natural Science Foundation of China (61871188), Guangzhou city science and technology research projects (201902020008).

REFERENCES

1. Itti, L.: Automatic foveation for video compression using a neurobiological model of visual attention. IEEE Trans. Image Process. 13(10), 1304–1318 (2004)
2. Donoser, M., et al.: Saliency driven total variation segmentation. In: IEEE 12th International Conference on Computer Vision, Kyoto (2009)
3. Rutishauser, U., et al.: Is bottom-up attention useful for object recognition? In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Washington, DC (2004)
4. Chang, K.Y., et al.: Fusing generic objectness and visual saliency for salient object detection. In: International Conference on Computer Vision, Barcelona (2011)
5. Jiang, P., et al.: Salient region detection by ufo: Uniqueness, focusness and objectness. In: Proceedings of the IEEE international conference on computer vision, Sydney (2013)
6. Huang, F., et al.: Salient object detection via multiple instance learning. IEEE Trans. Image Process. 26(4), 1911–1922 (2017)
7. Wang, L., et al.: Learning to detect salient objects with image-level supervision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu (2017)
8. Zeng, Y., et al.: Multi-source weak supervision for saliency detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach (2019)
9. Hou, Q., et al.: Deeply supervised salient object detection with short connections. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu (2017)
10. Zhang, P., et al.: Amulet: Aggregating multi-level convolutional features for salient object detection. In: Proceedings of the IEEE International Conference on Computer Vision, Venice (2017)
11. Zhang, P., et al.: Learning uncertain convolutional features for accurate saliency detection. In: Proceedings of the IEEE International Conference on Computer Vision, Venice (2017)
12. Alexe, B., Deselaers, T., Ferrari, V.: Measuring the objectness of image windows. IEEE Trans. Pattern Anal. Mach. Intell. 34(11), 2189–2202 (2012)
13. Zitnick, C.L., Dollár, P.: Edge boxes: Locating object proposals from edges. In: European conference on computer vision, Zurich (2014)
14. Dai, J., et al.: R-FCN: Object detection via region-based fully convolutional networks. In: Proceedings of the 30th International Conference on Neural Information Processing Systems, Barcelona (2016)
15. Li, X., et al.: Saliency detection via dense and sparse reconstruction. In: Proceedings of the IEEE International Conference on Computer Vision, Sydney (2013)
16. Yang, C., et al.: Saliency detection via graph-based manifold ranking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland (2013)
17. Cheng, M.M., et al.: Global contrast based salient region detection. IEEE Trans. Pattern Anal. Mach. Intell. 37(3), 569–582 (2014)
18. Zhu, W., et al.: Saliency optimization from robust background detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus (2014)
19. Achanta, R., et al.: Frequency-tuned salient region detection. In: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) 2009, Miami (2009)
20. Rother, C., Kolmogorov, V., Blake, A.: Grabcut: Interactive foreground extraction using iterated graph cuts. In: ACM Transactions on Graphics (TOG), vol. 23, pp. 309–314.ACM, New York (2004)
21. Jiang, H., et al.: Automatic salient object segmentation based on context and shape prior. In: Proceedings of the British Machine Vision Conference, University of Dundee, Scotland (2011)
22. Achanta, R., et al.: Slic superpixels compared to state-of-the-art superpixel methods. IEEE Trans. Pattern Anal. Mach. Intell. 34(11), 2274–2282 (2012)
23. Boykov, Y., Kolmogorov, V.: An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. IEEE Trans. Pattern Anal. Mach. Intell. 26(9), 1124–1137 (2004)
24. Wang, Z., et al.: Object instance annotation with deep extreme level set evolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach (2019)
25. Zhou, Z., et al.: Prior distribution-based statistical active contour model. Multimedia Tools and Applications 78(26), 35813–35833 (2019)
26. Wu, H., Appia, V., Yezei, A.: Numerical conditioning problems and solutions for nonparametric IID statistical active contours. IEEE Trans. Pattern Anal. Mach. Intell. 35(6), 1298–1311 (2012)
27. Kim, J., et al.: A nonparametric statistical method for image segmentation using information theory and curve evolution. IEEE Trans. Image Process. 14(10), 1486–1502 (2005)
28. Michailovich, O., Rathi, Y., Tannenbaum, A.: Image segmentation using active contours driven by the bhattacharyya gradient flow. IEEE Trans. Image Process. 16(11), 2787–2801 (2007)
29. Zhao, H.K., et al.: A variational level set method for multiphase image segmentation. IEEE Trans. Image Process. 16(11), 2787–2801 (2007)
30. Parzen, E.: On estimation of a probability density function and mode. The Annals of Mathematical Statistics 33(3), 1065–1076 (1962)
31. Otsu, N.: A threshold selection method from gray-level histograms. IEEE Trans. Syst. Man Cybern. 9(1), 62–66 (1979)
32. Tang, M., et al.: Grabcut in one cut. In: Proceedings of the IEEE International Conference on Computer Vision, Sydney (2013)
33. Chan, T.F., Vese, L.A.: Active contours without edges. IEEE Trans. Image Process. 10(2), 266–277 (2001)
34. He, K., et al.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas (2016)
35. Lin, T.Y., et al.: Microsoft COCO: Common objects in context. In: European Conference on Computer Vision, Zurich (2014)
36. Alpert, S., et al.: Image segmentation by probabilistic bottom-up aggregation and cue integration. IEEE Trans. Pattern Anal. Mach. Intell. 34(2), 315–327 (2011)
37. Li, Y., et al.: The secrets of salient object segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus (2014)
38. Yan, Q., et al.: Hierarchical saliency detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland (2013)
39. Cheng, M.M., et al.: Salentshape: group saliency in image collections. The Visual Computer 30(4), 443–453 (2014)
40. Jiang, H., et al.: Salient object detection: A discriminative regional feature integration approach. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland (2013)
41. Zhang, J., et al.: Minimum barrier salient object detection at 80 FPS. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago (2015)
42. Borji, A., et al.: Salient object detection: A benchmark. IEEE Trans. Image Process. 24(12), 5706–5722 (2015)
43. Abadi, M., et al.: TensorFlow: A system for large-scale machine learning. In: 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16), Savannah (2016)