Electronic Research Archive

http://www.aimspress.com/journal/era

ERA, 32(3): 1624–1645.
DOI: 10.3934/era.2024074
Received: 04 November 2023
Revised: 15 January 2024
Accepted: 16 January 2024
Published: 19 February 2024

Research article

Weakly supervised salient object detection via bounding-box annotation and SAM model

Xiangquan Liu and Xiaoming Huang*

Computer School, Beijing Information Science and Technology University, Beijing 100192, China

* **Correspondence:** Email: huangxm18@bistu.edu.cn.

Abstract: Salient object detection (SOD) aims to detect the most attractive region in an image. Fully supervised SOD based on deep learning usually needs a large amount of data with human annotation. Researchers have gradually focused on the SOD task using weakly supervised annotation such as category, scribble, and bounding-box, while these existing weakly supervised methods achieve limited performance and demonstrate a huge performance gap with fully supervised methods. In this work, we proposed one novel two-stage weakly supervised method based on bounding-box annotation and the recent large visual model Segment Anything (SAM). In the first stage, we regarded the bounding-box annotation as the box prompt of SAM to generate initial labels and proposed object completeness check and object inversion check to exclude low quality labels, then we selected reliable pseudo labels for the training initial SOD model. In the second stage, we used the initial SOD model to predict the saliency map of excluded images and adopted SAM with the everything mode to generate segmentation candidates, then we fused the saliency map and segmentation candidates to predict pseudo labels. Finally we used all reliable pseudo labels generated in the two stages to train one refined SOD model. We also designed a simple but effective SOD model, which can capture rich global context information. Performance evaluation on four public datasets showed that the proposed method significantly outperforms other weakly supervised methods and also achieves comparable performance with fully supervised methods.

Keywords: salient object detection; weakly supervised; segment anything; bounding-box annotation; deep learning

1. Introduction

Salient object detection (SOD) aims to detect regions of an image that can attract human attention. It has important applications in various vision tasks, such as object detection [1], image retrieval [2], and object tracking [3]. Traditional methods detect salient object through handcrafted features [4–8],

which usually achieve low performance.

The emergence of deep convolutional neural networks in recent years has greatly improved the performance of SOD. Fully supervised methods [9–13] have shown satisfactory results while needing expensive pixel-level annotation.

Due to the fully supervised deep learning algorithm relying on pixel-wise annotation, researchers have gradually focused on the SOD task using weakly supervised annotation. Most previous weakly supervised SOD methods focus on using image-level category [14–17] and scribble [18, 19] as weak supervision, which makes it difficult to accurately locate objects and sometimes miss some salient parts. To address the above problems, Zhang et al. [18] used bounding-box [20, 21] annotation, adopted the traditional unsupervised SOD methods and iterative refinement to obtain pseudo labels, and trained a SOD neural network. Wang et al. [22] considered the GrabCut algorithm [23] on the bounding-box annotation to generate pseudo labels, which is regarded as supervision to train the SOD model. Both these bounding-box annotation based methods [18, 22] achieve limited performance and demonstrate a huge performance gap with fully supervised methods. For example, compared with the fully supervised method multi-guidance SOD model (Mguid-Net) [13], the recent bounding-box annotation based weakly supervised method GC [22] shows higher mean absolute error (MAE) by 67.57% and lower maximum f-measure (F_β) by 9.64% on the DUTS-TE [15] benchmark.

In this paper, we propose one weakly supervised SOD method based on the bounding-box annotation and the recent large visual model Segment Anything (SAM) [24]. Given one training dataset and its bounding-box annotations, we first regard the bounding-box annotations as box prompt for the SAM model to generate segmentation as SOD initial pseudo labels. Although SAM is one powerful segmentation model, it still may segment incomplete object or incorrectly segment the background region as object. To conquer the influence of undesired results, object completeness and object inversion check is proposed to select reliable pseudo labels, then these reliable pseudo labels are regarded as supervision to train one initial SOD network. Excluding the images with unreliable pseudo labels from the initial model training is helpful to achieve higher performance, but also reduces the diversity of training data. To address the impact of dataset reduction and enrich the training dataset, we adopt the initial SOD network and SAM with the everything mode to generate more reliable pseudo labels, then train one refined SOD model.

The experimental results show that we achieve significantly better performance than other weakly supervised methods and also achieve comparable performance with fully supervised methods.

The main contributions of this work include: 1) We propose one method to generate initial pseudo labels, then carry object incomplete and object inversion check to select reliable pseudo labels for the training initial SOD model. 2) We propose one method combining the initial SOD model and SAM with the everything mode to predict more reliable pseudo labels and enrich the training dataset. 3) We propose one simple yet effective SOD network, which can capture rich global context information. 4) The proposed method significantly outperforms other weakly supervised methods and also achieves comparable performance with fully supervised methods.

2. Related work

In this section, we first review the previous SOD methods with weak supervision including image-level category, scribble, and bounding box annotation, then introduce the recent large visual

model SAM.

2.1. Weakly supervised SOD

To achieve a trade-off between labeling efficiency and performance, several weakly supervised methods have been proposed to detect salient object.

Due to existing large-scale classification datasets, image-level category based methods have received more attention. Wang et al. [15] first proposed to perform SOD with image-level category labels and design a foreground inference network to predict saliency. Li et al. [14] proposed to combine a coarse salient object activation map from the classification network and saliency maps generated from unsupervised methods as pixel-level annotation to train fully convolutional networks. Piao et al. [16] presented a multi-filter directive network including a saliency network as well as multiple directive filters, which is designed to extract and filter more accurate saliency cues from the noisy pseudo labels. Piao et al. [17] designed a noise-robust adversarial learning framework and a noise-sensitive training strategy to conquer noise pseudo labels generated from category.

The scribble annotation on foreground and background also can be regarded as weak supervision for SOD. Zhang et al. [18] designed one model based on scribble annotation and edge detection, then proposed one edge-structure-aware module as a supplement of scribble annotation. Yu et al. [19] proposed a local coherence loss to propagate the scribble labels to unlabeled regions based on image features and pixel distance.

Weakly supervised SOD with bounding-box annotation has received wide attention in the past two years. Liu et al. [25] used bounding-box annotation and the traditional methods to generate initial pseudo labels, then iteratively refined the initial pseudo labels by learning a multitask map refinement network with bounding-box annotation finally, they trained one salient object detector supervised by refined pseudo labels. Wang et al. [22] considered the GrabCut algorithm [23] on the bounding-box annotation and some adjustment steps to generate pseudo labels, which is regarded as supervision to train the SOD model.

Some researchers consider combining different forms of weak supervision. Zeng et al. [26] designed a multisource weak supervision framework to integrate category and caption labels.

Although the above weakly SOD methods were proposed in recent years, these existing weakly supervised methods achieve limited performance and demonstrate a huge performance gap with fully supervised methods. In this work, we propose one new weakly method and aim to achieve comparable performance with fully supervised methods.

2.2. SAM

With the development of deep learning [27–29], many deep learning-based image segmentation methods [30–32] have achieved significant results, but they usually rely on predefined conditions. Recently, SAM [24] was introduced to address challenges such as diverse object shapes, sizes, and complex background conditions. The technique aims to achieve efficient and accurate segmentation of arbitrary targets, regardless of target shape, size, and background conditions.

SAM can segment the corresponding mask based on prompts such as the given foreground or background points, bounding-box, mask, and text. The implementation of the above function consists of three steps: 1) extracting the image embedding, which is a time-consuming process but only needs

to be calculated once for each image, 2) encoding the user input prompt by the prompt encoder, and 3) decoding the image embedding and prompt embedding by the mask decoder, extracting the interested segmentation, and outputting the corresponding predicted masks and the intersection over union (IoU) score of each mask.

SAM is widely used in computer vision tasks. Chen et al. [33] combined image-level annotation and SAM for semantic segmentation. Yamagiwa et al. [34] proposed a novel zero-shot edge detection model with SAM. Zhang et al. [35] introduced a training-free object segmentation approach for SAM by one-shot tuning on a pair of an image and a mask. Chen et al. [36] incorporated domain-specific information or visual prompts into the segmentation network and significantly elevated the performance of SAM. Zhao et al. [37] proposed a speed-up alternative method for SAM with comparable performance.

In weakly supervised SOD with bounding-box annotation, generation pixel-level segmentation with bounding-box annotation is the most important step. We can regard the bounding-box annotation as a box prompt input for the SAM model to generate pixel-level segmentation.

3. Proposed method

The main flowchart of our method includes two stages, which is shown in Figure 1. In the first stage, given one training dataset and its bounding-box annotations, we regard the bounding-box annotations as box prompt for the SAM model to generate initial pseudo labels, then carry two label selection strategies to select reliable pseudo labels as supervision to train one initial SOD network. Excluding the images with low quality pseudo labels from the initial model training is a double-edged sword, which helps the model achieve higher performance but also reduces the diversity of training data. The second stage of our method aims to generate reliable pseudo labels for these excluded images and enrich the training dataset to train a refined SOD model. Specifically, we use the initial SOD network to predict the saliency of excluded images and adopt SAM with the everything mode to generate segmentation candidates, then fuse the saliency and segmentation candidates to predict pseudo labels. Finally we train one refined SOD model using all reliable labels generated in the two stages.

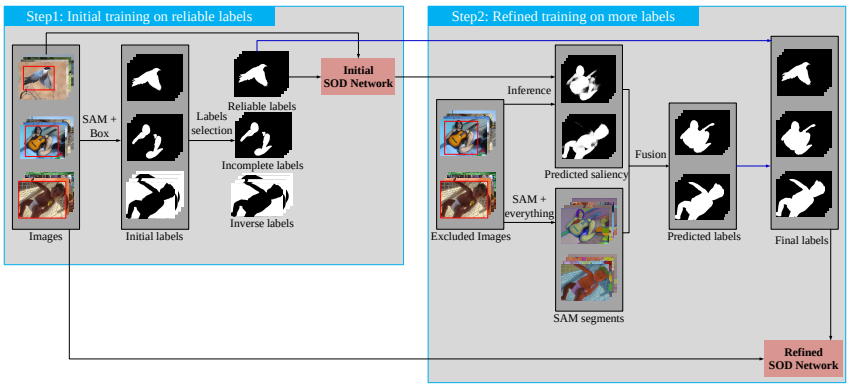

Figure 1. The main framework of the proposed method. In the first step, reliable initial labels are selected to train the initial SOD network and low quality labels are excluded. In the second step, some excluded labels are predicted for training the refined SOD network.

3.1. Bounding-box annotation

In this work, we use bounding-box annotation as weak supervision with two principles: 1) Each bounding-box is a minimum rectangular box that contains at least one salient object, and 2) the overlapped objects would be merged into one larger region and annotated by one bounding-box. Some examples of bounding-box annotation are shown in Figure 2.

Figure 2. Examples of bounding-box annotation.

3.2. Initial pseudo labels generation with SAM

According to the definition of bounding-box annotation, the outside of the bounding-box is the background and the inside of the bounding-box contains salient object while lacking pixel level location. Generation pixel level segmentation with bounding-box annotation is the key step of this task. The existing method [25] considered traditional SOD methods and the constraint of the bounding box to compute saliency, followed by saliency adjustment and refinement. However, these traditional SOD methods have poor generalization ability. In [22], authors proposed one bounding-box based segmentation via the GrabCut algorithm [23], which mainly considers low level image feature and leads to limited performance.

Recently, one new large visual segmentation model SAM [24] was proposed. SAM supports bounding-boxes as segmentation prompts. In this weakly supervised SOD task, the bounding-box annotation in each training image can be used as a box prompt input for the SAM model, and the SAM segmentation result can be regarded as an initial pseudo label for SOD. Figure 3 shows four examples of the initial pseudo label generated by the SAM model, where (a) is the input image and bounding-box annotation, (b) is manual annotated ground-truth, and (c) is the segmentation result of SAM with box prompt. The left three columns show satisfactory results, while the right three columns show undesired results.

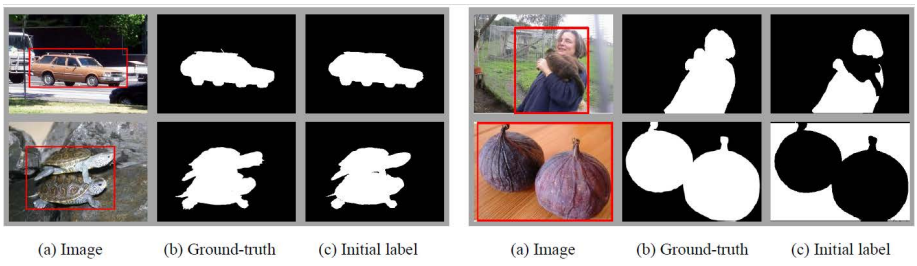

Figure 3. Some examples of SAM segmentation with box prompt. The left three columns show satisfactory results, while the right three columns show undesired results.

3.3. Initial pseudo label selection

Although the SAM model achieves satisfactory results in many training images, it still shows obvious deficiencies in some cases. Some object segmentation results are incomplete, and some background regions are segmented as object. Simply regarding all initial pseudo labels to train salient object detector will lead to limited performance. To conquer these problems, we propose object completeness and object inversion check to exclude undesired pseudo labels while reserving high quality labels.

Object completeness check. Given the input image and bounding-box annotation, which is shown in Figure 4(a), the ground truth is shown in Figure 4(b), and the SAM with box mode may generate an incomplete segmentation result. Two segmentation examples are shown in Figure 4(c), where the top and bottom demonstrate incomplete and complete segmentation results, respectively. To remove the incomplete result from training, we compute the mean IoU (mIoU) to measure the object completeness. Specifically, given one image and its bounding-box annotations B_i (i = 1..n), the segmentation result of SAM contains m regions and their minimum external rectangle can be denoted as C_j (j = 1..m). For each bounding-box B_i, we calculate its IoU with each segmentation rectangle C_j and select the maximum value, then calculate the mean IoU of all bounding-boxes:

$$IoU(B_i) = \max_{j=1..m} \left(\frac{|B_i \cap C_j|}{|B_i \cup C_j|}\right) \tag{3.1}$$

$$mIoU = \frac{1}{n}\sum_{i=1}^{n} IoU(B_i) \tag{3.2}$$

where | | means the area function. The $mIoU$ indicates the completeness of an initial pseudo label. In Figure 4(d), the top result shows $mIoU$ with value 0.596 and the bottom result shows $mIoU$ with value 0.971. For each image, its initial pseudo label with $mIoU$ is less than one threshold th_{mIoU}, which can be regarded as an incomplete label.

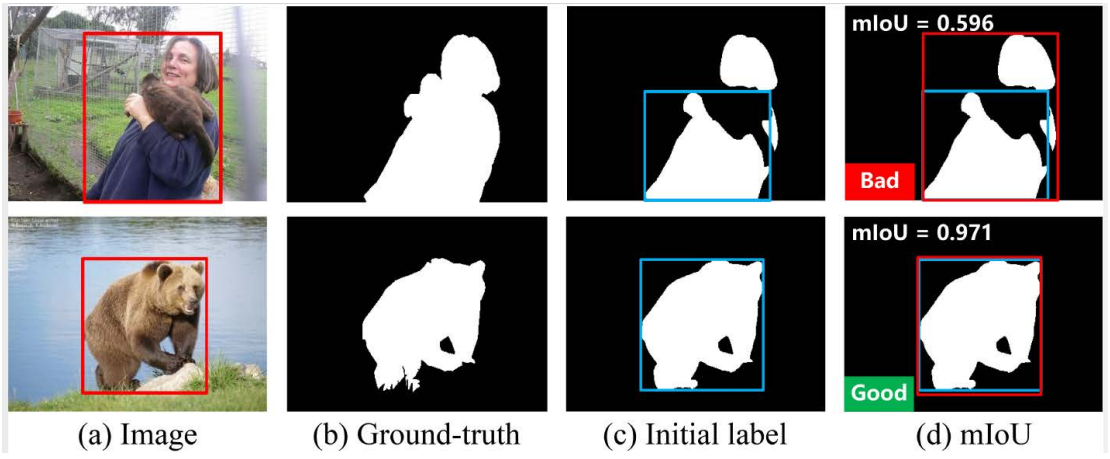

Figure 4. Examples of object completeness check. The top and bottom rows show incomplete and complete object segmentation result, respectively.

Object inversion check. When salient object occupies most of the image, SAM sometimes ignores the foreground region and incorrectly segments the background region as object. One example is shown in the top row of Figure 5. To conquer this problem, we first perform morphological dilatation filtering on the initial pseudo label, then compute the ratio of object pixels touching the boundary of the image as follows:

$$Bnd = \frac{E_t + E_b + E_l + E_r}{2 \times (H + W)} \tag{3.3}$$

where H and W are the height and width of the image, respectively, and E_t, E_b, E_l, and E_r are the numbers of pixels touched with top, bottom, left, and right boundary of the image. The Bnd measures the object inversion of the initial label. In Figure 5(d), the top result shows Bnd with value 0.923 and the bottom result shows Bnd with value 0.066. For each image, its initial pseudo label with Bnd larger than one threshold th_{Bnd} can be regarded as an inverse label.

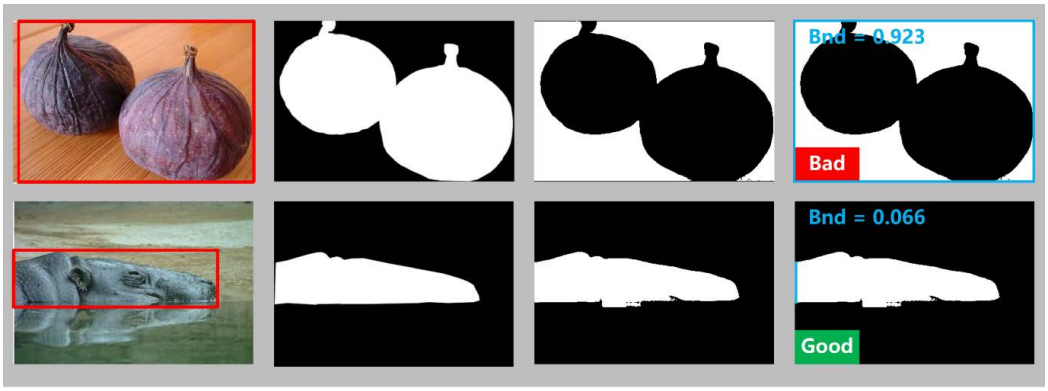

Figure 5. Examples of object inversion check. The top and bottom rows show inverse and correct object segmentation results, respectively.

3.4. Training initial SOD network

In the aforementioned section, given one training dataset, we adopt SAM with box mode to generate initial pseudo labels, which is followed by label selection to exclude undesired pseudo labels while reserving high quality pseudo labels. The reliable pseudo labels can be regarded as supervision to train one initial SOD network.

In this paper, inspired by [38], we propose one simple yet effective SOD network, which is depicted in Figure 6. The proposed SOD model contains three components: an encoder, an atrous spatial pyramid pooling (ASPP) module, and an attention mechanism based decoder. The encoder is based on the ResNet101 [39] to extract low-level to high-level visual features and reduce the resolution of feature maps. The encoder consists of a head-convolution and four residual layers denoted as $Layer_i$ (i $\in$ 1, 2, 3, 4) . In these four residual layers, the number of residual learning bottlenecks are 3, 4, 23, 3, the strides are set as 1, 2, 2, 1, and the dilations are set as 1, 1, 1, 2, respectively. The resolution of each residual layer feature is 1/4, 1/8, 1/16, 1/16 of the image resolution, respectively.

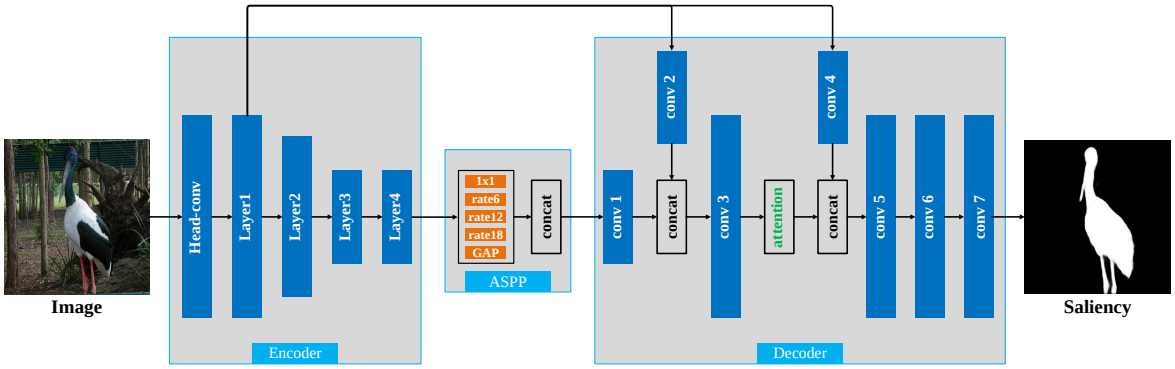

Figure 6. Network architecture of our method. The concat is concatenation operation among channel axis, and attention mechanism includes a channel attention and spatial attention operation.

The ASPP module performs multiscale feature extraction and fusion via dilated convolutions, which allows the extracted features to have a larger receptive field while preserving the resolution of the feature maps. In the ASPP module, the input feature pass through five parallel layers including a 1×1 convolution, three 3×3 convolutions with dilation rate of 12, 24, and 36, respectively, and a global average pooling layer, then the outputs of five parallel layers are concatenated to one high-level feature.

The decoder aims to predict a high-resolution saliency map with object details by fusion high-level features and low-level features. Specifically, the high-level output of the ASPP module is followed by 1×1 convolution ($conv1$) to yield feature F_1. Simultaneously, the low-level feature from $Layer_1$ is dimension reduced by 1×1 convolution ($conv2$) to yield 64-channels feature F_2. The high-level feature F_1 is concatenated with low-level feature F_2 and followed by a 3×3 convolution ($conv3$), which outputs 256-channels feature F_3.

To assign higher weights on potential salient regions and eliminate redundant noise, the feature F_3 is fed into one attention mechanism module, which contains a sequential channel attention operation and a spatial attention operation, inspired by [40]. Simultaneously, the $Layer_1$ feature is dimension reduced to 64-channels feature by 1×1 convolution ($conv4$) and concatenated with the outputs of the attention mechanism module to recover more object details:

$$F_4 = concat((Attn(F_3), conv4(Layer_1))) \tag{3.4}$$

where $concat$ is the concatenation operation among channel axis, $Attn$ denotes the attention mechanism, and $conv4$ denotes one convolution operation.

Finally, the concatenated feature F_4 passes through two 3×3 convolutional layers ($conv5$, $conv6$), one 1×1 convolutional layer ($conv7$), and one upsampling and sigmoid operations to predict saliency. The cross-entropy loss is selected as the loss function:

$$Loss = -\frac{1}{N}\sum_{i=1}^{N}\{PG_i \cdot log(S_i) + (1 - PG_i) \cdot log(1 - S_i)\} \tag{3.5}$$

where N means batch size, PG means the pseudo label, and S means the network predicted saliency.

3.5. Excluded pseudo labels prediction

In the training of the initial SOD model, some images with incomplete or inverse pseudo labels are excluded from the training dataset for better performance. This operation also reduces the diversity of training data and degrades the generalization of the model. In this section, we propose one method to generate reliable pseudo labels for these excluded images.

Given one excluded image I with bounding-box annotations B_i ($i = 1..n$), the desired pseudo labels need two requirements: salient and with accurate boundary. For the former condition, we use the trained initial SOD network to predict the image saliency S, which indicates salient region but lacks accurate object boundary. For the latter condition, we adopt SAM with the everything mode to generate all possible segmentation candidates R_i ($i = 1..k$), which can provide accurate object boundary. For each segmentation candidate R_i, in order to determine whether it is salient object or not, two factors are considered:

(1) The position relative to bounding-box annotations. We denote all pixels inside bounding-box annotations as M, then calculate the overlapped proportion between segmentation candidate R_i and M:

$$Rpos = \frac{|R_i \cap M|}{|R_i|} \tag{3.6}$$

where $|\ |$ means the area function.

(2) The mean saliency of segmentation candidate. We binarize the predicted saliency S as S_{bin}, then calculate pixel-level intersection ratio between segmentation candidate R_i and binaried saliency S_{bin}:

$$Rsal = \frac{|R_i \cap S_{bin}|}{|R_i|} \tag{3.7}$$

If both the above two metrics exceed threshold th_{Rpos} and th_{Rsal}, respectively, the segmentation R_i can be regarded as one salient region. All salient regions in the image I together form the predicted pseudo label.

One example of excluded pseudo labels prediction is shown in Figure 7. Given the input image and annotation, which is shown in Figure 7(a), its ground truth is shown in Figure 7(b), its initial pseudo label generated by SAM with box mode is incomplete object and is shown in Figure 7(c), its saliency map predicted by the initial SOD network and binarization result is shown in Figure 7(d) and (e), and SAM with the everything mode generates all possible segmentation candidates with accurate object boundary, which is shown in Figure 7(f). The segmentation candidates inside bounding-box annotation and having high saliency are selected to form predicted pseudo label, which is shown in Figure 7(g).

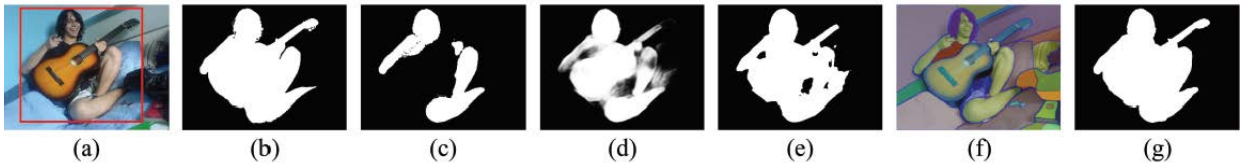

Figure 7. The diagram of the excluded pseudo label prediction. (a) input image and bounding-box annotation, (b) ground truth, (c) initial pseudo label generated by SAM with box mode, (d) saliency map predicted by initial SOD network, (e) binarized saliency map, (f) segmentation candidates generated by SAM with everything mode, and (g) predicted pseudo label.

The excluded pseudo label prediction can generate satisfactory results in most cases, but it inevitably failed in some cases. The object completeness check is applied on these excluded pseudo label prediction results once again to select reliable pseudo labels.

The described procedure can be represented using pseudo-code and is shown in Algorithm 1. All reliable pseudo labels generated in this stage and initial stage are regarded as supervision to train the refined SOD model.

Algorithm 1 Excluded pseudo labels prediction

Input: bounding-box annotations B_i ($i = 1..n$), segmentation candidates R_i ($i = 1..k$) generated by SAM everything mode, saliency map S predicted by initial SOD network

Output: pseudo-label mask PG and quality

$PG = []$
$S_{bin} =$ binarization(S)
$M =$ all pixels inside bounding-box annotations
for each segmentation candidate R_i **do**
    calculate the position relative to bounding-box annotation $Rpos$ using Eq 3.6
    calculate the mean saliency $Rsal$ using Eq 3.7
    if $Rpos > th_{Rpos}$ and $Rsal > th_{Rsal}$ **then**
        $PG = PG + R_i$
    end if
end for
$PG = PG \cap M$
calculate $mIoU$ between PG and bounding-box annotations using Eqs 3.1 and 3.2
if $mIoU > th_{mIoU}$ **then**
    return PG, *reliable*
else
    return PG, *unreliable*
end if

4. Experiments

In this section, to prove the effectiveness of the proposed method, we present experimental results including implementation details, evaluation metrics, performance comparison, and ablation study.

4.1. Implementation details

This work uses bounding-box annotation as the prompt of SAM to generate an initial label, then carries object completeness and object inversion check to select reliable labels with threshold parameter th_{mIoU} and th_{Bnd}, which is selected as 0.9 and 0.3, respectively. In the excluded pseudo label prediction stage, we use the initial trained SOD model to predict saliency followed by binarization with threshold 0.25 and we use SAM with the everything mode to generate all possible segmentation candidates, then select salient region with threshold parameter th_{Rpos} and th_{Rsal}, which is selected as 0.95 and 0.3, respectively. The choice of important parameters will be discussed in the ablation studies.

Our proposed SOD network uses ResNet101, which was pretrained on ImageNet [41] as encoder backbone. For the decoders, the convolutional layers weights are initialized by normal distribution with 0.01 standard deviation and zero mean value. Each 3×3 convolutional layer is followed by a relu and a batch norm layer and the 1×1 convolutional layer *conv*2 and *conv*4 are followed by a batch norm layer only. We use the stochastic gradient descent to train the SOD network 40 epochs with an initial learning rate of 2×10^{-8}, a weight decay of 5×10^{-4}, and a momentum of 0.9. The learning rate is decreased by 10% at 10 epochs and batch size is set as 10. To predict the pseudo label of excluded images in the first stage, we use the initial trained SOD network to predict their saliency and fuse SAM segmentation with the everything mode. For predicting better saliency with the initial SOD network, the image feed to training initial SOD network is resized to 512×512. In training the final SOD network, following fully supervised method stacked cross refinement network (SCRN) [9], we resize the input image to 352×352 and scaled resolution with [0.75, 1, 1.25] times to augment data.

4.2. Datasets

Train data: Following the weakly supervised method based on the bounding-box annotation [25], we take DUTS-TR [15] as the training set, which contains 10553 images. Although the training set already has pixel-level annotation, we do not use pixel-level annotation.

Test data: To verify the performance of the proposed method, tests were conducted on four public datasets: DUTS-TE [15], extended complex scene saliency dataset (ECSSD) [42], DUT-OMRON [8], and HKU-IS [43]. The DUTS-TE dataset contains 5019 test images, which contain important scenes for saliency detection. The ECSSD dataset consists of 1000 images with different sizes and multiple targets. The DUT-OMRON dataset contains 5168 images, and the dataset scenes are complex and contain one or multiple objects. The HKU-IS dataset contains 4447 images with high-quality pixel-wise annotations.

4.3. Evaluation metrics

Maximum F-measure(F_β): The performance measurement computed by the weighted harmonic of the precision and recall is below:

$$F_\beta = \frac{(1+\beta^2) \times P \times R}{\beta^2 \times P + R} \tag{4.1}$$

where β^2 is set to 0.3 to raise more importance on precision.

MAE: The difference between the saliency map S and the ground-truth G can be calculated as below:

$$MAE = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} |S(x,y) - G(x,y)| \tag{4.2}$$

4.4. Performance comparison

We compare our method against some recent state of the art methods, including fully supervised models progressive attention guided recurrent network (PAGR) [44], gated bi-directional message passing module (BMPM) [45], detect globally and refine locally (DGRL) [46], pixel-wise contextual attention network (PiCANet) [47], SCRN [9], visual saliency transformer (VST) [10], PoolNet-R+ [11], integrity cognition network (ICON-R) [12], Mguid-Net [13], weakly supervised models alternate saliency map optimization (AMSO) [14], weakly supervised saliency (WSS) [15], NWS [26], saliency bounding boxes (SBB) [25], GC [22], weakly-supervised SOD via scribble annotations (WSSA) [18], multi-filter directive network (MFNet) [16], structure-consistent weakly supervised SOD (SCWSSOD) [19], noise-sensitive adversarial learning (NSAL) [17], and unsupervised models texture-guided saliency distilling (TSD) [48], uncertainty mining network (UMNet) [49], unsupervised domain adaptive SOD (UDASOD) [50], and encoder-bigbigan (E-BigBiGAN) [51]. For fair comparison, we use either the implementations or the saliency maps provided by the authors.

Quantitative Comparison: Table 1 shows quantitative comparison between our work and nine fully supervised, nine weakly supervised, and four unsupervised methods. The up (down) arrow indicates that the larger (smaller) value is better. Compared with unsupervised and weakly supervised models, we significantly show better performance on all datasets and every evaluation metric. Compared with fully supervised models PAGR [44], BMPM [45], DGRL [46], and PiCANet [47], the proposed method also achieves better performance. Compared with fully supervised models SCRN [9], VST [10], PoolNet-R+ [11], ICON-R [12], and Mguid-Net [13], our work shows only a slightly worse, even comparable performance, while these methods need human labeling pixel-wise annotation.

We also trained our proposed SOD network with manual ground-truth of the training dataset DUTS-TR [15]. The parameters and training setting are the same as training the final SOD model in weakly mode. In manual ground-truth, although the object and background are usually presented as 1 and 0, we find that it is not binarized on the object boundary. Each training image is annotated by several people, and their annotations show little difference in object boundary. The final ground-truth is the mean of all annotations. Since our pseudo-label is binarized, for fair comparison, we trained our proposed SOD network with manual ground-truth and their binarization, which are denoted as Ours_GT and Ours_GT* in Table 1. Compared with Ours_GT*, our weakly supervised method shows comparable performance except for a slightly worse performance on challenging the DUT-OMRON dataset, and Ours_GT shows larger MAE, mainly due to lacking binarization on object boundary of ground-truth.

Qualitative Comparison: Figure 8 shows a few saliency maps of the evaluated methods. The visual comparison shows that our method performs well and is robust, and can adapt to complex or small object scenes.

Table 1. Quantitative comparison of our method with other methods. The up (down) arrow indicates that the larger (smaller) value is better, the bold representation means the best result.

Methods	Supervision	Input size	ECSSD		DUTS-TE		HKU-IS		DUT-OMRON	
			$F_\beta \uparrow$	MAE↓	$F_\beta \uparrow$	MAE↓	$F_\beta \uparrow$	MAE↓	$F_\beta \uparrow$	MAE↓
PAGR [44]	Fully	353×353	0.927	0.061	0.855	0.056	0.918	0.048	0.711	0.071
BMPM [45]	Fully	256×256	0.928	0.044	0.850	0.049	0.920	0.038	0.775	0.063
DGRL [46]	Fully	384×384	0.925	0.043	0.834	0.051	0.914	0.037	0.779	0.063
PiCANet [47]	Fully	224×224	0.935	0.047	0.860	0.051	0.919	0.043	0.803	0.065
SCRN [9]	Fully	352×352	0.950	0.038	0.888	0.040	0.934	0.034	0.812	0.056
VST [10]	Fully	224×224	**0.951**	0.033	0.890	**0.037**	**0.942**	**0.029**	0.825	0.058
PoolNet-R+ [11]	Fully	300×400	0.949	0.040	**0.894**	0.039	0.941	0.034	**0.831**	0.056
ICON-R [12]	Fully	384×384	0.950	**0.032**	0.892	**0.037**	0.939	**0.029**	0.825	0.057
Mguid-Net [13]	Fully	352×352	0.946	0.035	0.892	**0.037**	0.938	0.031	0.805	0.056
Ours_GT	Fully	352×352	0.944	0.042	0.893	0.038	0.939	0.033	0.811	**0.055**
Ours_GT*	Fully	352×352	0.948	0.035	0.890	**0.037**	**0.942**	**0.029**	0.812	**0.055**
AMSO [14]	Weakly	328×328	0.810	0.114	0.625	0.123	0.821	0.091	0.633	0.100
WSS [15]	Weakly	256×256	0.828	0.105	0.657	0.106	0.821	0.081	0.611	0.111
NWS [26]	Weakly	256×256	0.846	0.096	0.704	0.097	0.823	0.086	0.619	0.109
SBB [25]	Weakly	256×256	0.878	0.072	0.775	0.073	0.869	0.057	0.751	0.075
GC [22]	Weakly	300×400	0.894	0.062	0.806	0.062	0.880	0.052	0.791	0.065
WSSA [18]	Weakly	256×256	0.884	0.061	0.780	0.062	0.874	0.047	0.740	0.068
MFNet [16]	Weakly	256×256	0.873	0.084	0.763	0.079	0.875	0.058	0.685	0.098
SCWSSOD [19]	Weakly	320×320	0.915	0.049	0.844	0.049	0.909	0.038	0.783	0.060
NSAL [17]	Weakly	256×256	0.878	0.078	0.781	0.073	0.882	0.051	0.715	0.088
Ours	Weakly	352×352	**0.945**	**0.035**	**0.889**	**0.036**	**0.932**	**0.031**	**0.793**	**0.055**
TSD [48]	Unsupervised	320×320	0.926	0.044	0.843	**0.047**	0.914	0.037	**0.790**	0.061
UMNet [49]	Unsupervised	320×320	0.903	0.063	0.798	0.067	0.907	0.041	0.787	0.063
UDASOD [50]	Unsupervised	352×352	**0.929**	**0.043**	**0.850**	0.050	**0.923**	**0.035**	0.778	**0.059**
E-BigBiGAN [51]	Unsupervised	512×512	0.825	0.162	0.668	0.195	0.804	0.155	0.607	0.232

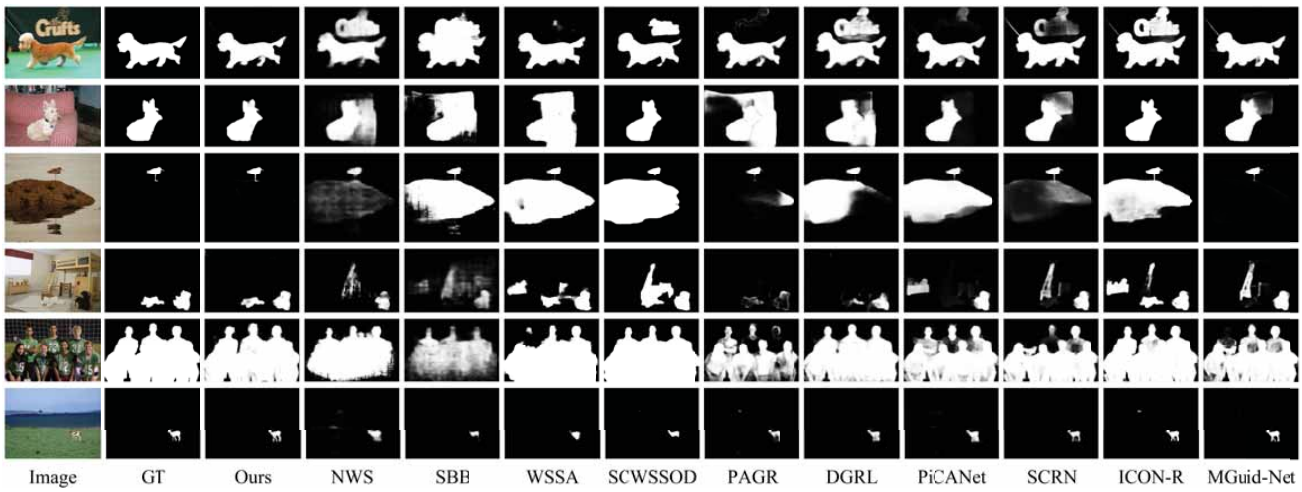

Figure 8. Saliency quality comparison between our work with recent methods.

4.5. Ablation study

4.5.1. Approach effectiveness analysis

In this work, we first adopt the bounding-box annotations to generate initial pseudo labels and select reliable pseudo labels to train one initial SOD model, then predict more reliable pseudo labels to enrich the training dataset and learn the refined SOD model. For fair comparison, we follow SCRN [9] and use multiscale augmented training data to train the final SOD model. To verify the effectiveness of each step, we also train one baseline SOD model with all initial pseudo labels. The performance of these SOD models is shown in Table 2. Compared with the baseline model, the initial SOD model achieves significant performance improvement, especially the MAE metric reduced by 13.04%. This result proved that the removal of low-quality labels for model training is effective. The refined SOD model further improves performance and the MAE metric is reduced by 5%, which mainly benefited from the excluded pseudo label prediction. The final SOD model demonstrates the best performance.

Table 2. The performance comparison of SOD models with different training settings on DUTS-TE dataset. The up (down) arrow indicates that the larger (smaller) value is better.

Method	Training Setting	Image Num	$F_\beta \uparrow$	MAE$\downarrow$
Baseline	all initial labels	10553	0.854	0.046
Initial SOD	reliable initial labels	8452	0.869	0.040
Refined SOD	reliable initial label +reliable predicted label	9847	0.875	0.038
Final SOD	reliable initial label +reliable predicted label +multiscale training	9847	0.889	0.036

4.5.2. Initial reliable labels selection

To generate initial reliable labels from SAM with box mode, we carry object completeness and object inversion check to select labels. For each image, its initial pseudo label with object completeness measure $mIoU$ less than threshold th_{mIoU} or object inversion measure Bnd larger than one threshold th_{Bnd} will be removed from the initial SOD model training. The label selection is one key step of our method. In order to analyze the parameter influence of the label selection, we use different threshold parameter settings to select reliable labels and train the initial SOD model, then evaluate the model performance. The experimental results are shown in Table 3. The parameter value $\times$ means the selection parameter is invalid and turns off the selection operation.

We first turn off the two label selection operations as baseline, which is shown in the top row of Table 3. We set object inversion parameter th_{Bnd} with 0.9, 0.6, 0.5, 0.3, and 0.15 and set object completeness parameter th_{mIoU} with 0.3, 0.6, 0.7, 0.9, and 0.95. The smaller parameter th_{Bnd} and the larger parameter th_{mIoU} means more strict label selection condition, which can generate more reliable pseudo labels and achieve performance improvement. On the other hand, the excessively strict label selection condition also decreases the number and diversity of training data, which leads to performance drop. The parameter th_{Bnd} with 0.3 and the parameter th_{mIoU} with 0.9 shows the best performance, which is selected as the best parameter.

Table 3. The label selection parameter analysis on DUTS-TE dataset. The up (down) arrow indicates that the larger (smaller) value is better, the bold representation means the best result.

th_{Bnd}	th_{mIoU}	Initialed SOD		
		Image Num	$F_\beta \uparrow$	MAE $\downarrow$
×	×	10553	0.854	0.046
0.90	0.30	10328	0.858	0.045
0.60	0.60	9803	0.865	0.043
0.50	0.70	9555	0.863	0.043
0.30	0.90	8452	**0.869**	**0.040**
0.15	0.95	6848	0.861	0.042

Table 4. The single th_{Bnd} parameter analysis on DUTS-TE dataset. The up (down) arrow indicates that the larger (smaller) value is better, the bold representation means the best result.

th_{Bnd}	Initialed SOD		
	Image Num	$F_\beta \uparrow$	MAE $\downarrow$
0.90	10494	0.858	0.047
0.60	10473	0.858	0.047
0.50	10460	0.858	0.047
0.30	10393	**0.860**	**0.045**
0.15	9997	0.855	0.046

Table 5. The single th_{mIoU} parameter analysis on DUTS-TE dataset. The up (down) arrow indicates that the larger (smaller) value is better, the bold representation means the best result.

th_{mIoU}	Initialed SOD		
	Image Num	$F_\beta \uparrow$	MAE $\downarrow$
0.30	10393	0.859	0.047
0.60	9963	0.859	0.044
0.70	9709	**0.865**	**0.041**
0.90	8612	0.861	0.042
0.95	7234	0.861	0.042

We also explored the effect of only using object inversion parameter th_{Bnd} or object completeness parameter th_{mIoU}, which is shown in Tables 4 and 5, respectively. Parameters of th_{Bnd} and th_{mIoU} settings remain the same as in Table 3. The experimental results conclusion is similar to Table 3. The smaller parameter th_{Bnd} and the larger parameter th_{mIoU} means more strict label selection condition, which can generate more reliable pseudo labels and achieve performance improvement. On the other hand, the excessively strict label selection condition also decreases the number and diversity of training data, which leads to a performance drop. When only using object inversion parameter th_{Bnd} to select reliable labels, th_{Bnd} with 0.30 still shows the best performance in Table 4, but demonstrates a slightly

worse performance than the performance in Table 3 when th_{Bnd} with 0.30 and th_{mIoU} with 0.90. This phenomenon proves that object inversion check and object completeness check are complementation operations to select reliable labels. One similar conclusion is also presented in Table 5.

4.5.3. Excluded label prediction

To predict the pseudo label of excluded images in the first stage, we use the initial SOD network to predict their saliency and adopt SAM with the everything mode to generate segmentation candidates. For each segmentation candidate, we calculate its position measure $Rpos$ and saliency measure $Rsal$. If both $Rsal$ and $Rpos$ exceed threshold th_{Rpos} and th_{Rsal} ,the segmentation candidate will be considered as a salient region. We set parameter th_{Rpos} with 0.60, 0.90, 0.95, 0.97, 0.99 and set th_{Rsal} with 0.1, 0.2, 0.3, 0.5, 0.7, then train the refined SOD model and evaluate the performance, which is shown in Table 6. The model performs the best performance when the parameter th_{Rpos} is set to 0.95 and th_{Rsal} is set to 0.3.

Table 6. The excluded pseudo label prediction parameter analysis on DUTS-TE dataset. The up (down) arrow indicates that the larger (smaller) value is better and the bold representation means the best result.

th_{Rpos}	th_{Rsal}	Refined SOD		
		Image Num	$F_\beta \uparrow$	MAE $\downarrow$
0.60	0.1	9990	0.872	0.040
0.90	0.2	9900	0.873	0.039
0.95	0.3	9847	**0.875**	**0.038**
0.97	0.5	9750	0.873	0.039
0.99	0.7	9620	0.871	0.039

Table 7. The single th_{Rpos} parameter analysis on DUTS-TE dataset. The up (down) arrow indicates that the larger (smaller) value is better and the bold representation means the best result.

th_{Rpos}	Refined SOD		
	Image Num	$F_\beta \uparrow$	MAE $\downarrow$
0.60	10313	0.859	0.051
0.90	10261	0.868	0.044
0.95	10244	0.868	0.044
0.97	10232	0.863	0.045
0.99	10214	**0.874**	**0.041**

In addition, we further discuss the result of only using position measure parameter th_{Rpos} and saliency measure parameter th_{Rsal}, which is shown in Tables 7 and 8, respectively. Parameters of th_{Rpos} and th_{Rsal} settings remain the same as in Table 6. When only using position measure parameter th_{Rpos} to predict reliable labels of excluded images, higher parameter th_{Rpos} means more strict condition, which can generate more reliable pseudo labels and achieve performance improvement.

For the parameter th_{Rpos}, even when selected as the very large value 0.99, there are still 10214 images for training the refined SOD model and maintaining the diversity of training data. When only using saliency measure parameter th_{Rsal} to predict reliable labels of excluded images, higher parameter th_{Rsal} means more strict condition, which can generate more reliable pseudo labels and achieve performance improvement. On the other hand, the excessively strict condition also decreases the number and diversity of training data, which leads to performance drop. The parameter th_{Rsal} with 0.3 still shows the best performance in Table 8, but demonstrates a slightly worse performance than the performance in Table 6 when th_{Rsal} with 0.3 and th_{Rpos} with 0.95. This phenomenon proves that position measure and saliency measure are complementation measures to select reliable labels.

Table 8. The single th_{Rsal} parameter analysis on DUTS-TE dataset. The up (down) arrow indicates that the larger (smaller) value is better and the bold representation means the best result.

th_{Rsal}	Refined SOD		
	Image Num	$F_\beta \uparrow$	MAE $\downarrow$
0.1	10017	0.872	0.039
0.2	9952	0.870	0.039
0.3	9907	**0.873**	**0.038**
0.5	9800	0.872	0.039
0.7	9672	0.868	0.041

5. Conclusions

Fully supervised SOD based on deep learning usually needs a large amount of data with human annotation. The weakly supervised method based on category, scribble, and bounding-box consumes low annotation cost while achieving limited performance and demonstrating a huge performance gap with fully supervised methods. Research on the weakly supervised method with low annotation cost but achieving comparable performance with fully supervised methods is still challenging for the SOD task.

In this paper, we propose one weakly supervised SOD using bounding-box annotation with two stages. In the first stage, we use the bounding-box annotation as the box prompt of SAM to generate initial pseudo labels, then carry object completeness and object inversion check to select reliable pseudo labels for training the initial SOD model. In the second stage, we use the initial SOD model and SAM with the everything mode to predict more reliable pseudo labels, then use the reliable pseudo labels generated in the two stages as supervision to train one refined SOD model. Experiments show that we achieve significantly better performance than other weakly supervised methods, and also achieve comparable performance with other fully supervised methods.

However, the proposed method contains two stages and is difficult to achieve global optimization. In the future, we will consider one end-to-end method to solve this problem. In addition, we will consider further reducing the annotation cost with unsupervised methods such as k-means clustering [52] and contrastive learning. We also will consider applying this method to related fields such as SOD on optical remote sensing images [53–55].

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This work was supported by R&D Program of Beijing Municipal Education Commission (KM202011232014).

Conflict of interest

The authors declare there is no conflicts of interest.

References

1. J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2016), 779–788. https://doi.org/10.1109/CVPR.2016.91
2. X. Yang, X. Qian, Y. Xue, Scalable mobile image retrieval by exploring contextual saliency, *IEEE Trans. Image Process.*, **24** (2015), 1709–1721. https://doi.org/10.1109/TIP.2015.2411433
3. Y. Su, Q. Zhao, L. Zhao, D. Gu, Abrupt motion tracking using a visual saliency embedded particle filter, *Pattern Recognit.*, **47** (2014), 1826–1834. https://doi.org/10.1016/j.patcog.2013.11.028
4. X. Huang, Y. Zhang, 300-fps salient object detection via minimum directional contrast, *IEEE Trans. Image Process.*, **26** (2017). 4243–4254,https://doi.org/10.1109/TIP.2017.2710636
5. X. Huang, Y. Zhang, Water flow driven salient object detection at 180 fps, *Pattern Recognit.*, **76** (2018), 95–107. https://doi.org/10.1016/j.patcog.2017.10.027
6. X. Huang, Y. Zheng, J. Huang, Y. Zhang, 50 fps object-level saliency detection via maximally stable region, *IEEE Trans. Image Process.*, **29** (2020), 1384–1396. https://doi.org/10.1109/TIP.2019.2941663
7. M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, S. Hu, Global contrast based salient region detection, *IEEE Trans. Pattern Anal. Mach. Intell.*, **37** (2015), 569–582. https://doi.org/10.1109/TPAMI.2014.2345401
8. C. Yang, L. Zhang, H. Lu, X. Ruan, M. Yang, Saliency detection via graph-based manifold ranking, in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, (2013), 3166–3173. https://doi.org/10.1109/CVPR.2013.407
9. Z. Wu, L. Su, Q. Huang, Stacked cross refinement network for edge-aware salient object detection, in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, (2019), 7263–7272. https://doi.org/10.1109/ICCV.2019.00736
10. N. Liu, N. Zhang, K. Wan, L. Shao, J. Han, Visual saliency transformer, in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, (2021), 4722–4732. https://doi.org/10.1109/ICCV48922.2021.00468

11. J. Liu, Q. Hou, Z. Liu, M. Cheng, PoolNet+: Exploring the potential of pooling for salient object detection, *IEEE Trans. Pattern Anal. Mach. Intell.*, **45** (2023), 887–904. https://doi.org/10.1109/TPAMI.2021.3140168

12. M. Zhuge, D. Fan, N. Liu, D. Zhang, D. Xu, L. Shao, Salient object detection via integrity learning, *IEEE Trans. Pattern Anal. Mach. Intell.*, **45** (2023), 3738–3752. https://doi.org/10.1109/TPAMI.2022.3179526

13. S. Hui, Q. Guo, X. Geng, C. Zhang, Multi-guidance cnns for salient object detection, *ACM Trans. Multimedia Comput. Commun. Appl.*, **19** (2023), 1–19. https://doi.org/10.1145/3570507

14. G. Li, Y. Xie, L. Lin, Weakly supervised salient object detection using image labels, in *Thirty-Second AAAI Conference on Artificial Intelligence*, **32** (2018), 7024–7031. https://doi.org/10.1609/aaai.v32i1.12308

15. L. Wang, H. Lu, Y. Wang, M. Feng, D. Wang, B. Yin, et al., Learning to detect salient objects with image-level supervision, in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2017), 3796–3805. https://doi.org/10.1109/CVPR.2017.404

16. Y. Piao, J. Wang, M. Zhang, H. Lu, Mfnet: Multi-filter directive network for weakly supervised salient object detection, in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, (2021), 4116–4125. https://doi.org/10.1109/ICCV48922.2021.00410

17. Y. Piao, W. Wu, M. Zhang, Y. Jiang, H. Lu, Noise-sensitive adversarial learning for weakly supervised salient object detection, *IEEE Trans. Multimedia*, **25** (2023), 2888–2897. https://doi.org/10.1109/TMM.2022.3152567

18. J. Zhang, X. Yu, A. Li, P. Song, B. Liu, Y. Dai, Weakly-supervised salient object detection via scribble annotations, in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2020), 12543–12552. https://doi.org/10.1109/CVPR42600.2020.01256

19. S. Yu, B. Zhang, J. Xiao, E. G. Lim, Structure-consistent weakly supervised salient object detection with local saliency coherence, in *AAAI Conference on Artificial Intelligence*, (2021), 3234–3242. https://doi.org/10.1609/aaai.v35i4.16434

20. J. Dai, K. He, J. Sun, Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation, in *2015 IEEE International Conference on Computer Vision (ICCV)*, (2015), 1635–1643. https://doi.org/10.1109/ICCV.2015.191

21. A. Khoreva, R. Benenson, J. Hosang, M. Hein, B. Schiele, Simple does it: Weakly supervised instance and semantic segmentation, in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2017), 1665–1674. https://doi.org/10.1109/CVPR.2017.181

22. Q. Wang, X. Huang, Q. Tong, X. Liu, Weakly supervised salient object detection algorithm based on bounding box annotation, *J. Comput. Appl.*, **43** (2023), 1910–1918.

23. C. Rother, V. Kolmogorov, A. Blake, "GrabCut": Interactive foreground extraction using iterated graph cuts, *ACM Trans. Graphics*, **23** (2004), 309–314. https://doi.org/10.1145/1015706.1015720

24. A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, et al., Segment anything, preprint, arXiv:2304.02643.

25. Y. Liu, P. Wang, Y. Cao, Z. Liang, R. W. H. Lau, Weakly-supervised salient object detection with saliency bounding boxes, *IEEE Trans. Image Process.*, **30** (2021), 4423–4435. https://doi.org/10.1109/TIP.2021.3071691

26. Y. Zeng, Y. Zhuge, H. Lu, L. Zhang, M. Qian, Y. Yu, Multi-source weak supervision for saliency detection, in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2019), 6067–6076. https://doi.org/10.1109/CVPR.2019.00623

27. J. Lu, L. Pan, J. Deng, H. Chai, Z. Ren, Y. Shi, Deep learning for flight maneuver recognition: A survey, *Electron. Res. Arch.*, **31** (2023), 75–102. https://doi.org/10.3934/era.2023005

28. Z. Feng, K. Qi, B. Shi, H. Mei, Q. Zheng, H. Wei, Deep evidential learning in diffusion convolutional recurrent neural network, *Electron. Res. Arch.*, **31** (2023), 2252–2264. https://doi.org/10.3934/era.2023115

29. J. Wang, L. Zhang, S. Yang, S. Lian, P. Wang, L. Yu, et al., Optimized LSTM based on improved whale algorithm for surface subsidence deformation prediction, *Electron. Res. Arch.*, **31** (2023), 3435–3452. https://doi.org/10.3934/era.2023174

30. C. Swarup, K. U. Singh, A. Kumar, S. K. Pandey, N. varshney, T. Singh, Brain tumor detection using CNN, AlexNet & GoogLeNet ensembling learning approaches, *Electron. Res. Arch.*, **31** (2023), 2900–2924. https://doi.org/10.3934/era.2023146

31. R. Bi, L. Guo, B. Yang, J. Wang, C. Shi, 2.5D cascaded context-based network for liver and tumor segmentation from CT images, *Electron. Res. Arch.*, **31** (2023), 4324–4345. https://doi.org/10.3934/era.2023221

32. S. Kara, H. Ammar, F. Chabot, Q. Pham, Image segmentation-based unsupervised multiple objects discovery, in *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, (2023), 3276–3285. https://doi.org/10.1109/WACV56688.2023.00329

33. T. Chen, Z. Mai, R. Li, W. Chao, Segment anything model (SAM) enhanced pseudo labels for weakly supervised semantic segmentation, preprint, arXiv:2305.05803.

34. H. Yamagiwa, Y. Takase, H. Kambe, R. Nakamoto, Zero-shot edge detection with SCESAME: Spectral clustering-based ensemble for segment anything model estimation, preprint, arXiv:2308.13779.

35. R. Zhang, Z. Jiang, Z. Guo, S. Yan, J. Pan, X. Ma, et al., Personalize segment anything model with one shot, preprint, arXiv:2305.03048.

36. T. Chen, L. Zhu, C. Ding, R. Cao, Y. Wang, Z. Li, et al., SAM fails to segment anything?–SAM-adapter: Adapting SAM in underperformed scenes: Camouflage, shadow, and more, preprint, arXiv:2304.09148.

37. X. Zhao, W. Ding, Y. An, Y. Du, T. Yu, M. Li, et al., Fast segment anything, preprint, arXiv:2306.12156.

38. H. Li, G. Chen, G. Li, Y. Yu, Motion guided attention for video salient object detection, in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, (2019), 7273–7282. https://doi.org/10.1109/ICCV.2019.00737

39. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2016), 770–778. https://doi.org/10.1109/CVPR.2016.90
40. F. Deng, H. Feng, M. Liang, H. Wang, Y. Yang, Y. Gao, et al., FEANet: Feature-enhanced attention network for RGB-thermal real-time semantic segmentation, in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, (2021), 4467–4473. https://doi.org/10.1109/IROS51168.2021.9636084
41. J. Deng, W. Dong, R. Socher, L. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, (2009), 248–255. https://doi.org/10.1109/CVPR.2009.5206848
42. Q. Yan, L. Xu, J. Shi, J. Jia, Hierarchical saliency detection, in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, (2013), 1155–1162. https://doi.org/10.1109/CVPR.2013.153
43. G. Li, Y. Yu, Visual saliency detection based on multiscale deep CNN features, *IEEE Trans. Image Process.*, **25** (2016), 5012–5024. https://doi.org/10.1109/TIP.2016.2602079
44. X. Zhang, T. Wang, J. Qi, H. Lu, G. Wang, Progressive attention guided recurrent network for salient object detection, in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2018), 714–722. https://doi.org/10.1109/CVPR.2018.00081
45. L. Zhang, J. Dai, H. Lu, Y. He, G. Wang, A bi-directional message passing model for salient object detection, in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2018), 1741–1750. https://doi.org/10.1109/CVPR.2018.00187
46. T. Wang, L. Zhang, S. Wang, H. Lu, G. Yang, X. Ruan, et al., Detect globally, refine locally: A novel approach to saliency detection, in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2018), 3127–3135. https://doi.org/10.1109/CVPR.2018.00330
47. N. Liu, J. Han, M. Yang, PiCANet: Learning pixel-wise contextual attention for saliency detection, in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2018), 3089–3098. https://doi.org/10.1109/CVPR.2018.00326
48. H. Zhou, B. Qiao, L. Yang, J. Lai, X. Xie, Texture-guided saliency distilling for unsupervised salient object detection, in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2023), 7257–7267. http://doi.org/10.1109/cvpr52729.2023.00701
49. Y. Wang, W. Zhang, L. Wang, T. Liu, H. Lu, Multi-source uncertainty mining for deep unsupervised saliency detection, in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2022), 11717–11726. https://doi.org/10.1109/CVPR52688.2022.01143
50. P. Yan, Z. Wu, M. Liu, K. Zeng, L. Lin, G. Li, Unsupervised domain adaptive salient object detection through uncertainty-aware pseudo-label learning, in *AAAI Conference on Artificial Intelligence*, (2022), 3000–3008. https://doi.org/10.1609/aaai.v36i3.20206
51. A. Voynov, S. Morozov, A. Babenko, Object segmentation without labels with large-scale generative models, preprint, arXiv:2006.04988.
52. S. Jardim, J. António, C. Mora, Graphical image region extraction with k-means clustering and watershed, *J. Imaging*, **8** (2022), 163. https://doi.org/10.3390/jimaging8060163

53. G. Li, Z. Liu, D. Zeng, W. Lin, H. Ling, Adjacent context coordination network for salient object detection in optical remote sensing images, *IEEE Trans. Cybern.*, **53** (2023), 526–538. https://doi.org/10.1109/TCYB.2022.3162945

54. G. Li, Z. Liu, Z. Bai, W. Lin, H. Ling, Lightweight salient object detection in optical remote sensing images via feature correlation, *IEEE Trans. Geosci. Remote Sens.*, **60** (2022), 1–12. https://doi.org/10.1109/TGRS.2022.3145483

55. G. Li, Z. Bai, Z. Liu, X. Zhang, H. Ling, Salient object detection in optical remote sensing images driven by transformer, *IEEE Trans. Image Process.*, **32** (2023), 5257–5269. https://doi.org/10.1109/TIP.2023.3314285

AIMS Press