Mathematical Biosciences and Engineering

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

MBE, 20(12): 21359–21381.
DOI: 10.3934/mbe.2023945
Received: 02 October 2023
Revised: 20 November 2023
Accepted: 23 November 2023
Published: 01 December 2023

Research article

Weakly supervised salient object detection via image category annotation

Ruoqi Zhang[1], **Xiaoming Huang**[1,*] **and Qiang Zhu**[1,2]

[1] Computer School, Beijing Information Science and Technology University, Beijing 100192, China

[2] College of Computer Science and Technology, Zhejiang University, Hangzhou 310013, China

* **Correspondence:** Email: huangxm18@bistu.edu.cn, huangxm0556@163.com; Tel: +8613426005342.

Abstract: The rapid development of deep learning has made a great progress in salient object detection task. Fully supervised methods need a large number of pixel-level annotations. To avoid laborious and consuming annotation, weakly supervised methods consider low-cost annotations such as category, bounding-box, scribble, etc. Due to simple annotation and existing large-scale classification datasets, the category annotation based methods have received more attention while still suffering from inaccurate detection. In this work, we proposed one weakly supervised method with category annotation. First, we proposed one coarse object location network (COLN) to roughly locate the object of an image with category annotation. Second, we refined the coarse object location to generate pixel-level pseudo-labels and proposed one quality check strategy to select high quality pseudo labels. To this end, we studied COLN twice followed by refinement to obtain a pseudo-labels pair and calculated the consistency of pseudo-label pairs to select high quality labels. Third, we proposed one multi-decoder neural network (MDN) for saliency detection supervised by pseudo-label pairs. The loss of each decoder and between decoders are both considered. Last but not least, we proposed one pseudo-labels update strategy to iteratively optimize pseudo-labels and saliency detection models. Performance evaluation on four public datasets shows that our method outperforms other image category annotation based work.

Keywords: weakly supervised; salient object detection; saliency detection; image category annotation; deep learning

1. Introduction

Salient object detection aims to simulate the human visual system for detecting regions that are most attractive, which is widely used in many computer vision tasks such as image segmentation [1,2], defect detection [3,4], object tracking [5], etc. Traditional methods detect salient object through hand-crafted features [6,7], which usually achieve low performance.

In recent years, deep neural networks (DNN) have achieved great progress in many related fields [8–10]. DNN has also been widely used in salient object detection [11–15]. Fully supervised methods [16–18] have shown satisfactory results while requiring expensive pixel-level annotation. An experienced annotator often takes several minutes to label one image, and the expensive labeling cost makes it difficult to train a model using large-scale labeled datasets.

Due to fully supervised methods relying on expensive pixel-wise annotation, researchers have considered weakly supervised salient object detection. The weak supervision mainly includes bounding box labeling [19], image category labeling [20–22], scribble labeling [23], etc. In these weakly supervised methods, a category annotation based model can be trained with existing large-scale classification datasets (e.g., ImageNet). It can largely reduce the annotation cost while satisfying the requirement for massive data and attract the attention of many researchers.

Category label based weakly supervised methods usually first obtain the coarse location of the foreground via classification network. This step can be implemented with class activation mapping (CAM) [24], while CAM often only provides the most discriminative object part. Wang et al. [25] proposed one method to jointly learn image feature and foreground mask, then performed mask on image feature, which is followed by classification, the classification task guide, the mask and segment foreground. To transform a spatial-dependent masked image feature to a position-independent classification feature, they integrated global max pooling (GMP) and global average pooling (GAP), which needs high computational cost. Since GMP only focuses on the most discriminative object part and often fails to discover the full objects, we propose one coarse object location network (COLN) that only uses GAP. The COLN uses the image classification dataset with category labels to jointly train a foreground inference network (FIN) and an image classification network and generate a FIN map, which means coarse foreground location. Additionally, to conquer the shortcoming of GAP, which may lead to overestimated object areas, we propose one method based on classification accuracy analysis to generate the optimal FIN map that covers the object region and excludes the background region.

In category label based weakly supervised methods, coarse object locations need some refinement to generate pixel-level pseudo-labels which are used to train saliency detection networks. The noise in pseudo-label is a key factor for model performance. We propose one quality check strategy to select high quality pseudo-labels. To this end, we use a slightly different model parameter to learn COLN twice and to obtain one pair of pseudo-labels, when calculating the consistency of pseudo-label pairs to select high quality labels. The high quality pseudo-labels are more robust to the choice of model parameters. A slight adjustment on a model parameter leads to less influence on high quality pseudo-labels and more influence on low quality pseudo-labels. The consistency calculation on two pseudo-labels with a slightly different model parameter is helpful to select high quality pseudo-labels.

Learning one robust saliency detection model from pseudo-labels is also one important step of the weakly supervised method. For this purpose, we propose one multi-decoder neural network (MDN) for saliency detection supervised by a pseudo-label pairs. The MDN is designed to integrate different saliency information from pseudo-label pairs generated by COLN. Each decoder branch is responsible for one pseudo-label. The loss of each decoder and consistency between decoders are both considered. Additionally, we also propose one pseudo-labels update strategy to further optimize the pseudo-labels and improve performance of saliency detection.

Performance evaluation on four public datasets shows that the proposed method outperforms other image category supervision based work.

The main contributions of this work include:

- We propose one COLN to roughly locate the object of an image.
- We propose one novel quality check strategy to generate high quality pseudo labels for saliency detection training.
- We propose one MDN for salient object detection, which is supervised by high quality pseudo-label pairs simultaneously.
- We propose one pseudo-labels update strategy, which iteratively optimizes the pseudo-labels and saliency detection model.

2. Related work

Over the past few decades, researchers have developed many methods for salient object detection. In the early days, researchers [26–30] focused more on traditional features (e.g., color, texture, contrast) to detect the salient region. Deep learning has achieved amazing results in computer vision in recent years, and subsequently, various saliency detection methods based on deep learning have been proposed. These methods mainly include fully supervised and weakly supervised methods.

2.1. Fully supervised salient object detection

The fully supervised method uses manual pixel-level annotation as a supervision signal to learn one saliency detection model.

Some methods focused on fusion multi-scale features and enhancing image boundaries to improve detection results. Tang et al. [31] decomposed the task into a low-resolution saliency classification network which aims to identify explicit image regions, and a high-resolution refinement network to precisely refine the saliency values of pixels in uncertain regions. Ma et al. [32] aimed to shrink pairwise aggregated neighboring feature nodes layer by layer so that the aggregated features fuse valid details and semantics together, meanwhile discarding distracting information. Song et al. [33] introduced an implicit function to simulate the equilibrium state of the feature pyramid at infinite depth, and they also proposed a differentiable edge extractor that directly extracts edges from the saliency masks. Zhuge et al. [34] aimed to learn more complete salient objects, aggregate multi-scale features and enhance salient objects. Wu et al. [35] proposed a dynamic convolutional kernel size to capture objects with different sizes. Wu et al. [36] employed an extreme downsampling technique to effectively learn a global view of the whole image and constructed an elegant decoder for recovering object details. Ma et al. [37] designed an enhanced wider field of sensation framework, which allows the network to achieve very significant improvements when dealing with objects with scale variations.

Recent attention mechanisms [38–41], which focus on important regions of an image, are also widely used in salient object detection tasks. Liu et al. [42] developed a new unified model based on pure transformers, which also improves the performance for high-resolution images. Wang et al. [43] proposed curiosity-driven network and fragment attention to generate enhanced detail-rich saliency maps based on curiosity. Xie et al. [44] designed a framework extracting features from images at different sizes and resolutions, as well as used transformers and convolutional neural network (CNN) backbone independently to achieve promised results on high-resolution images. Fan et al. [45] proposed a high-quality dataset and a data enhancement strategy to train the model to adapt to each

different complex scene. Cheng et al. [46] proposed an extremely lightweight model that is about 0.2% parameters size of the current popular larger models and gains comparable performance. Tian et al. [47] designed a selective object saliency module and an object-context-object relation module to unify spatial attention and object-based attention for saliency ranking.

Although these fully supervised methods achieve satisfactory results, they require expensive and time-consuming pixel-level annotations.

2.2. Weakly supervised salient object detection

To reduce the cost of image annotation, and enable larger datasets to be applied to salient object detection tasks, weakly supervised methods have attracted the interest of researchers.

Image category is one widely considered weak supervision for salient object detection. Wang et al. [25] designed one method that can only use image category annotation. They first divided the image classification task into two subtasks: Image foreground inference and foreground classification. In the image foreground inference task, they used a foreground inference network to obtain one roughly salient foreground, which was refined into pseudo-labels for the network training. This work alleviates the cost of annotation and allows the use of existing large-scale training sets with image-level labels. Li et al. [20] proposed to use the combination of a coarse salient object activation map from the classification network and saliency maps generated from unsupervised methods as pixel-level annotation, and they developed a simple yet very effective algorithm to train fully convolutional networks for salient object detection supervised by these noisy annotations. The algorithm is based on alternately exploiting a graphical model and training a fully convolutional network for model updating. Piao et al. [21] proposed a multi-filter directive network including a saliency network, as well as multiple directive filters. The directive filter is designed to extract and filter more accurate saliency cues from the noisy pseudo labels. The multiple accurate cues from multiple directive filters are then simultaneously propagated to the saliency network with a multi-guidance loss. Piao et al. [22] observed pseudo-labels converted from image-level classification labels always containing noise information and designed a noise-robust adversarial learning framework and a noise-sensitive training strategy to mitigate this problem. Tian et al. [48,49] proposed a novel weakly supervised network with category and subitizing labels for salient instance detection problems.

However, compared to full supervision, image category labeling loses most of the detailed information. Therefore, other researchers have also considered other forms of weak annotation. Zhang et al. [23] trained one network using scribble labeling and designed an edge detection module and an edge-structure-aware module to complement the scribble annotation. Liu et al. [19] proposed a novel weakly supervised method by bounding boxes annotations. They first take the unsupervised methods to generate initial saliency maps and address the over/under prediction problems to obtain the initial pseudo-labels, then iteratively refine the initial labels by learning a multitask map refinement network with saliency bounding boxes. Liang et al. [50] also used bounding boxes annotations on salient object detection in light fields and proposed a fused attention module to utilize light field data from multi-view. Zheng et al. [51] introduced saliency subitizing as the weak supervision. They proposed a saliency subitizing module to generate the initial saliency masks using the subitizing information and a saliency updating module to iteratively refine the generated saliency masks. Liu et al. [52] introduced a label decoupling siamese network to more adequately use the scribble labels and the complementary relationship between salient objects and backgrounds. Zeng et al. [53, 54] introduced a unified frame-

work and learned image saliency from category labels, captions, web images, and untagged images, respectively. They designed a classification network and a title production network to predict object categories and generate titles, respectively, while highlighting potential foreground regions. Both networks are also encouraged to detect generally salient regions rather than task-specific regions. They then used predicted foreground regions to generate pseudo-labels to train a saliency detection network.

Although these weakly supervised methods achieved some progress, they still suffer from inaccurate detection of salient objects.

3. Methods

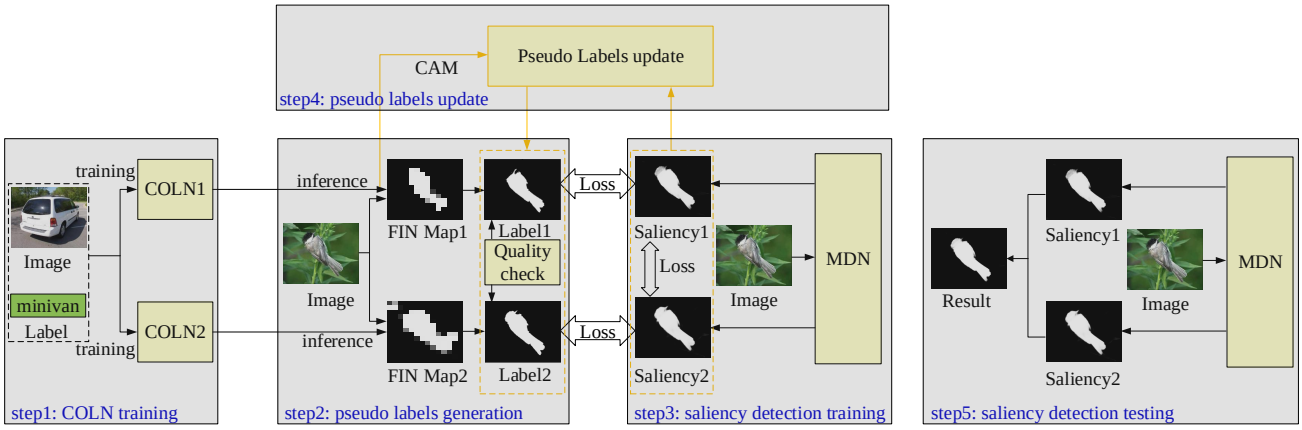


Figure 1. The framework of proposed method including five steps. 1) Training COLN on image classification datasets. 2) Generation high quality pseudo-labels on training dataset of salient object detection. 3) Training MDN for salient object detection, which is supervised by high quality pseudo-labels. 4) Pseudo-labels update, which further declines impact of pseudo-labels noise and learns more robust model. 5) Testing salient object detection.

The framework of proposed method is shown in Figure 1, which contains five steps:

1) Training COLN. We use the image classification dataset with category labels to train a FIN jointly with an image classification network and propose one method to generate the optimal FIN map, which covers the object region and excludes the background.

2) Generation high quality pixel-level pseudo-labels on training dataset of saliency detection. We perform COLN inference on training data to obtain coarse object location followed by GrabCut, then adopt one novel quality check strategy to generate high quality pseudo-labels pair.

3) Training MDN for salient object detection, which is supervised by a high quality pseudo-labels pair simultaneously. The MDN network consists of an encoder and two decoders, and each decoder and encoder forms a U-shaped structure.

4) Pseudo-labels update, which further declines impact of pseudo-labels noise and learns a more robust model.

5) Testing salient object detection. In the testing mode, the average of two saliency results predicted by MDN is regarded as the final saliency.

3.1. COLN training

The salient region is often represented by the foreground of an image in classification task. To segment the foreground of an image by a classification task, inspired by [25], one COLN is first proposed, as shown in Figure 2. Given one image with category label, the image is encoded with shared convolutional layers then fed forward through a full convolutional network (FCN) and an FIN to obtain the image feature map and the approximate foreground mask, respectively.

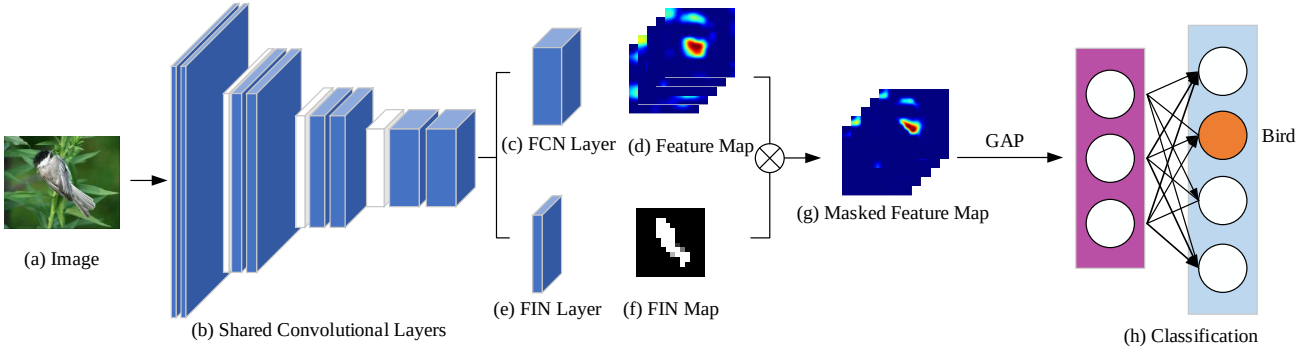


Figure 2. The pipeline of COLN. FIN generates the approximate foreground mask, which can be regarded as coarse object location.

Specifically, we adopt the shared convolutional layers on top of the 16-layer VGG [58] network, including 13 convolutional layers with rectified linear unit (ReLU) nonlinear interleaved layers and four max-pooling layers. FCN and FIN are two sibling sub-networks built on top of the shared layers. The FCN includes a convolutional layer, a batch normalization (BN) layer and a ReLU layer to generate feature map F with 512 channels. The FIN consists of a convolutional layer, a BN and a sigmoid layer to yield a saliency mask M with single channel, which is used to obtain a masked feature map as follows:

$$\hat{F}_k = F_k \odot M, \tag{3.1}$$

where F_k denotes the k-th channel of the feature map obtained by FCN and $\hat{F}_k$ is the k-th channel of the masked feature map $\hat{F}$. The $\odot$ means the element point multiplication.

One GAP layer is used to aggregate the masked feature map $\hat{F}$ into a 512-dimensional image-level feature, which is then passed through a fully connected layer and softmax operation to generate the category probabilities.

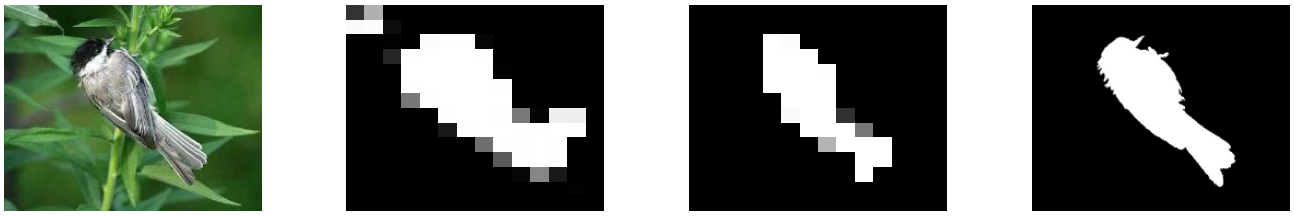
To prevent FIN from simply having high responses at all locations, a sparse regularization term is added to the loss function to penalize the high response of FIN on the background. Given a training set $\{X_i, l_i\}_{i=1}^N$ including N sample pairs (images X_i and label l_i), the loss function of this network consists of two parts as follows:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\left[\sum_{k\in l_i}\log\left(p_k\left(X_i\right)\right) + \sum_{k\notin l_i}\log\left(1-p_k\left(X_i\right)\right) - \lambda\left\|M\left(X_i\right)\right\|_1\right], \tag{3.2}$$

where the first and second terms are cross-entropy loss to measure classification accuracy, the third term is the L1 regularization on the saliency mask M predicted by FIN and $p_k(X_i)$ is the outputed probability of k-th category.

In loss function (3.2), λ is a predefined weight parameter to control the FIN map size. A larger λ will result in more constraint on the response area of FIN, which will tend to detect incomplete objects. In contrast, smaller λ reduces the constraint on the FIN response area, which allows FIN to overestimate the object and brings the risk of high response on background. One example of FIN map with different λ is shown in Figure 3. The optimal parameter λ is designed to get the optimal FIN map, which covers the object region and excludes the background.

In this work, we propose one method to find the optimal parameter λ over the entire dataset. We carefully analyze the relationship between parameter λ and the classification accuracy. The top-1 classification accuracy against parameter λ is shown in Figure 4. When λ value is very small, the constraint on the FIN response area is limited and the generated FIN contains both object and part of the background, which leads to low classification accuracy. As the parameter λ increases gradually, more penalty is carried on the FIN response area and the generated FIN mainly contains object and excludes the background, which leads to higher classification accuracy. However, if the parameter λ value is too large, increasing penalty on the FIN response area results in the generated FIN only containing incomplete objects and providing lower classification accuracy. Intuition, the optimal parameter λ, which is designed to get the optimal FIN map that covers the object region and excludes the background, obtains the highest classification accuracy. From Figure 4, the optimal parameter λ is set to 2.0×10^{-5}.



(a) Image (b) Lower λ (c) Higher λ (d) Ground Truth

Figure 3. Given one input image in (a), the generated FIN map with different parameter λ is shown in (b) and (c).

We train COLN classification network on one image classification dataset with category labels, then use the COLN network to perform inference on one image without any annotation and obtain its FIN map as the coarse object location.

Our proposed COLN is related to the network proposed by [25], while it exists two main differences:

1) **Different global feature aggregation.** In weak supervision with category-level labels, some form of global pooling is required to transform a spatial related image feature to a position-independent category feature. Both GMP and GAP have been intensively investigated in the literature. GMP only focuses on the most discriminative object part and often fails to discover the full objects. Zhou et al. [24] holds that GMP is not a suitable solution for the segmentation problem. In contrast, GAP encourages the network to have the same response at all positions and leads to overestimated object areas. The previous work WSS [25] uses a global smooth pooling (GSP) method, which integrates GAP and GMP, while the integration of two feature aggregation methods increases computational cost. Aiming at the deficiency of the method, we only use GAP in the proposed method. To avoid the shortcoming of GAP, which tends to overestimate object areas, we propose one method to calculate the optimal parameter λ for the FIN map which covers the object region and excludes the background.

2) **Double COLN training strategy.** Locating the foreground of an image by a classification task inevitably exists noise and leads to a significant performance drop. To conquer this problem, we train COLN twice with a slightly different parameter λ and follow with a quality check to select high quality results (presented in Section 3.2), while the network for FIN generation in WSS [25] is only trained once and is difficult to eliminate noise.

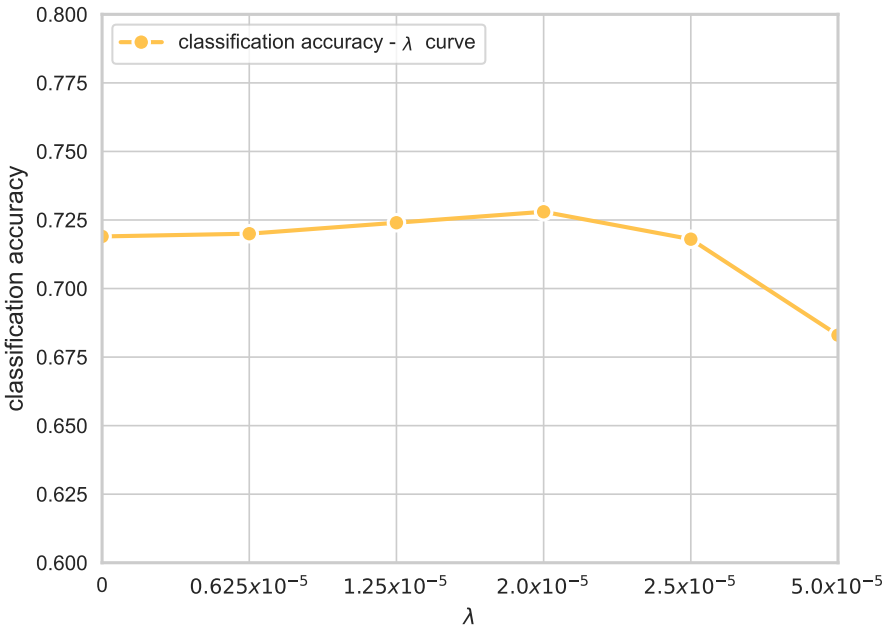


Figure 4. The top-1 classification accuracy with respect to different λ. As λ increases, the classification accuracy slightly increases. When λ is too large, the classification accuracy drops dramatically.

3.2. High quality pixel-level pseudo labels generation

We trained the COLN classification network on one image classification dataset with category labels in Section 3.1, then used COLN to perform inference on the saliency training dataset DUTS-TR [25] and obtained the FIN map of each image as coarse object location. The generated coarse object location via the FIN map only has a 16×16 resolution, which not only lacks detail structure but also exists noise. We further adopt the unsupervised image segmentation algorithm GrabCut [55, 56] to generate a pixel-level pseudo-label, and propose one quality check strategy to select high quality pseudo-labels.

GrabCut [55] is one efficient interactive object segmentation method. In this work, the coarse FIN map can be regarded as one interaction condition for GrabCut. Given one image with the FIN map, we regard the mean map value as one threshold. For each pixel, the map value larger (smaller) than the threshold will be initialized with *probable foreground* (*probable background*). Based on this initial label, GrabCut algorithm estimates the color distribution of the foreground and background via a Gaussian mixture model, constructs a graph over the entire image and applies a graph cut [56] optimization to achieve the segmentation result as pseudo-labels. Two segmentation examples are shown in step two of Figure 1.

The generated pseudo-labels with the FIN map and GrabCut still have noise and reduce the saliency

detection performance. One solution is to select high quality pseudo-labels and remove low quality parts. To this end, we learn COLN is twice followed by GrabCut refinement, the parameter λ is selected as the optimal value 2.0×10^{-5} and one approximately optimal value 2.5×10^{-5} to obtain one pair pseudo-labels, respectively. The slightly different selection on λ leads to these pair results as not entirely equal. For each image I, it generates one pair pseudo-label Y_1 and Y_2. We find that Y_1 and Y_2 are often consistent if they both are high quality results and Y_1 and Y_2 are often largely different if they contain low quality results. We calculate intersection over union (IOU) of the pseudo-label pairs as their consistency measure:

$$IOU = \frac{|Y_1 \cap Y_2|}{|Y_1 \cup Y_2|}, \tag{3.3}$$

where || means the area of the pseudo-label. Some examples are shown in Figure 5. In the top two rows, the pseudo-labels pair contains low quality results and leads to a low IOU. In the bottom two rows, the high quality pseudo-labels pair leads to high IOU.

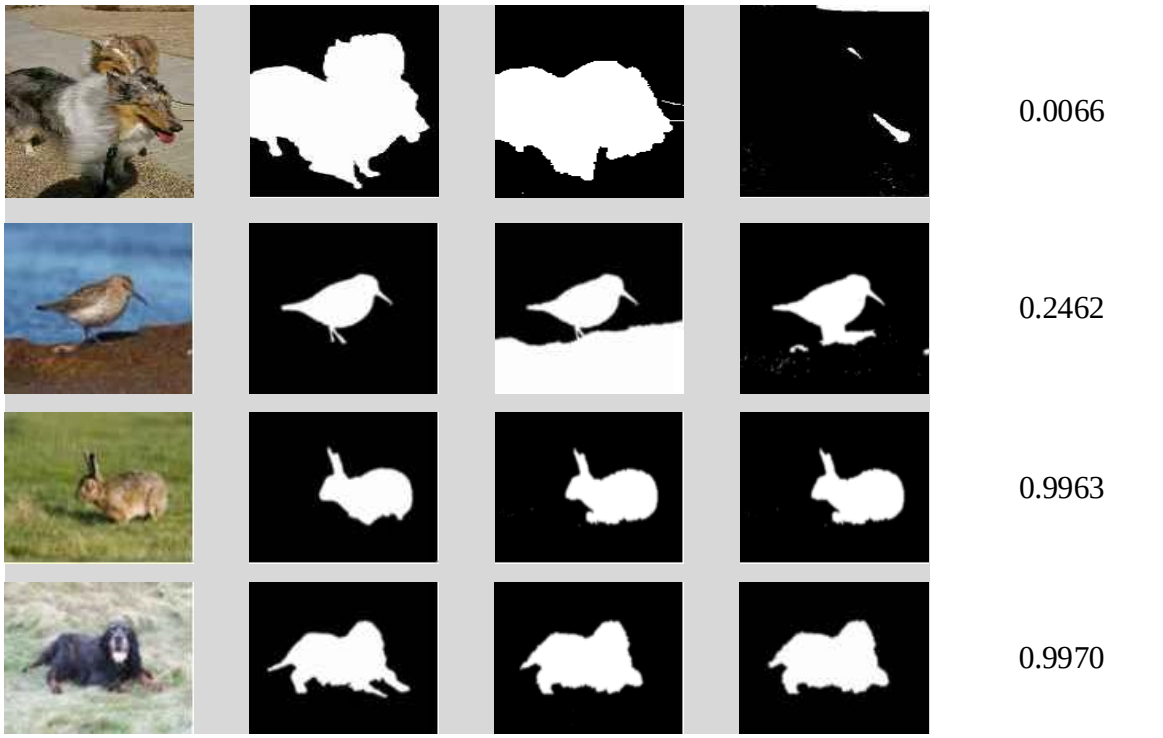


Figure 5. The IOU examples of a pseudo-label pair. Given one input image in (a), the generated pseudo-label pairs are shown in (c) and (d) and their IOU value is shown in (e).

To further analyze the relationship between IOU and quality pseudo-labels, for the pseudo-label pairs of each image in saliency training dataset DUTS-TR [25], we calculate their IOU and average root mean absolute error (MAE) to indicate absolute difference between pseudo-label pairs and ground-truths. The pseudo-labels are divided into five groups according to IOU (0.0–0.2, 0.2–0.4, etc.) and the image number and average MAE of each group is shown in Table 1. From this statistical result, in general we can find the pseudo-labels with larger IOU ddemonstrating lower MAE, which means

higher quality; the conclusion is consistent with Figure 5. Additionally, we can find the pseudo-labels with *IOU* value in 0.6–0.8, and 0.8–1.0 show comparable MAE and occupy about 87% volume of the dataset. On the other hand, the pseudo-labels with *IOU* value less than 0.6 show significant deterioration in MAE.

Table 1. Statistical analysis of *IOU* and MAE on pseudo-labels.

IOU	Image Num	MAE
0.8–1.0	5871	0.098
0.6–0.8	3338	0.110
0.4–0.6	125	0.173
0.2–0.4	403	0.171
0.0–0.2	816	0.201

Based on the above discussion, the pseudo-label pairs with high *IOU* probably are high quality pseudo-labels, otherwise they contain low quality pseudo-labels. Inspired by this intuition, we compute the *IOU* of each pseudo-labels pair. The pseudo-labels with *IOU* larger than one threshold θ (e.g., 0.6) will be regarded as high quality pseudo-labels to train salient object detection model.

3.3. MDN for salient object detection

To implement robust saliency detection, we propose MDN, which is supervised by a high quality pseudo-labels pair simultaneously.

The proposed network consists of an encoder and two decoders, which is shown in Figure 6. The input image is first encoded by one shared encoder (e.g., VGGNet-16) and then passed through pyramid pooling module (PPM) [57] to capture important global information of the input image. Taking the VGGNet version of the shared encoder as an example, encoded feature maps corresponding to $C = \{C_2, C_3, C_4, C_5\}$ in the pyramid have downsampling rates of two, four, eight and 16 compared to the size of the input image, respectively. For each decoder branch, the top feature C_5 is first sent through the feature aggregation module (FAM) [57] to fuse the coarse and fine level semantic features, then upsampled by factor two, concatenated with upsampled PPM and feature C_4 and finally passed through a 3*3 convolution layer and get a fused feature. The operation on this fused feature is similar to C_5. This process will be repeated four times and predict one saliency map. The loss of MDN is defined as:

$$L = \sum_{k=1}^{2} L_{bce}\left(P_k, Y_k\right) + \delta \times L_{ss}\left(P_1, P_2\right), \tag{3.4}$$

where L_{bce} is the binary cross entropy loss (BCE-loss) on one decoder branch, L_{ss} is the similarity loss between two decoders and δ is weight parameter.

$$L_{bce}\left(P_k, Y_k\right) = -\frac{1}{n}\sum_{i}^{n}[y_{ki} * \log p_{ki} + (1 - y_{ki}) * \log\left(1 - p_{ki}\right)], \tag{3.5}$$

$$L_{ss}\left(P_1, P_2\right) = \frac{1}{n}\sum_{i}^{n}\left(p_{1i} - p_{2i}\right)^2, \tag{3.6}$$

where p_{ki} and y_{ki} represent the elements of the decoder predictions P_k and its pseudo-labels Y_k.

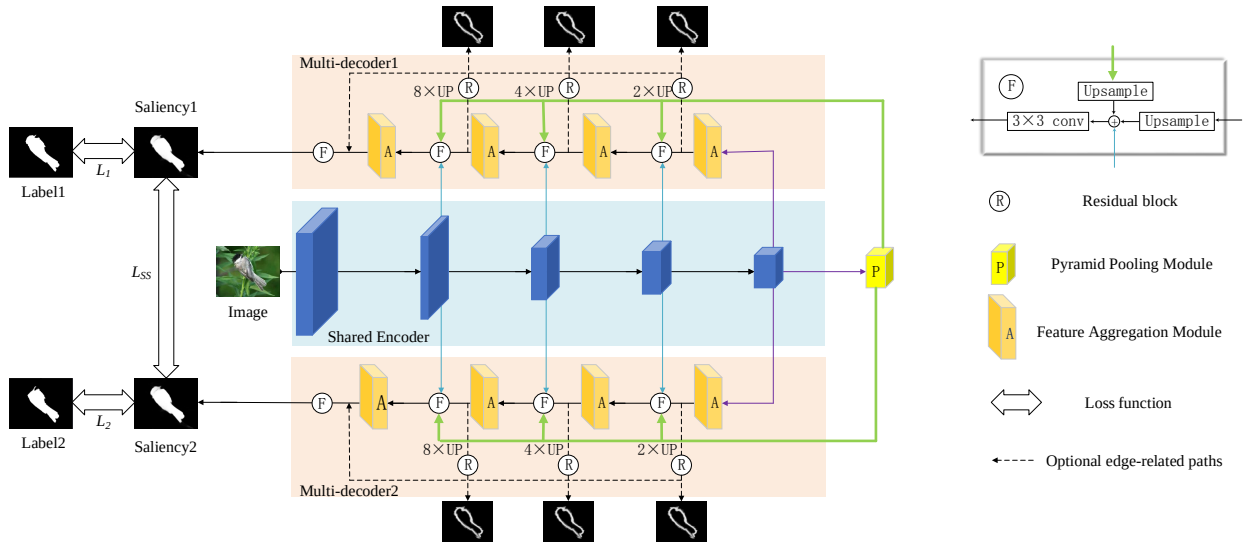


Figure 6. The proposed MDN contains a shared encoder and two decoders. Each decoder and encoder form a U-shaped structure. The top and bottom part for edge detection is optional.

Following the earlier work [57], we also jointly train salient object detection and edge detection to further improve performance. We add an extra edge prediction branch built upon each decoder branch to estimate the boundaries of the salient objects, which is shown on the top and bottom side of Figure 6. Based on the FAM [57], we add three residual blocks followed by a 16-channel 3×3 convolutional layer for feature compression and a one-channel 1×1 convolutional layer for edge prediction. We also concatenate these three 16-channel 3×3 convolutional layers and feed them to three consecutive 3×3 convolutional layers with 48 channels to transmit the captured edge information to the salient object detection decoder branch for detail enhancement.

3.4. Pseudo-labels update for iterative learning

The proposed dual decoder architecture adopts two high quality pseudo-labels for model learning. However, in this straightforward approach, the noise in pseudo-labels still may exist and lead to a decline of performance. To conquer the impact of noise, we propose one pseudo-labels update strategy to learn a more robust model.

As described in the previous section, we adopt the COLN classification network to perform inference on saliency training dataset DUTS-TR [25], followed by GrabCut and a high quality pseudo-labels selection strategy to get high quality pseudo-labels, then train an MDN saliency detection network. Given one training image, its average of pseudo-labels pair (Y_1 and Y_2) is denoted as Y_A and its average of predictions by MDN (P_1 and P_2) is denoted as P_A. The previous work WSS [25] found that in classification network, the average of the score map across all the channels followed by refinement can also be regarded as one saliency map, which is denoted as R_{CAM}. MAE () is the mean absolute difference between the two saliency maps. Rfn is the pixel-level refinement algorithm GrabCut.

We first compare the average pseudo-labels Y_A and the average predicted saliency P_A. Their MAE

is lower than one small threshold α, meaning they both are confident. We regard their average as one updated pseudo-label and regard the further result with Rfn refinement as another updated pseudo-label. If the MAE is higher than one large threshold β, it indicates that the pseudo-labels are probably wrong and should be discarded from the training dataset. In other cases, we regard R_{CAM} as reference, select one more reliable result from Y_A and P_A as one updated pseudo-label and regard the further result with Rfn refinement as another updated pseudo-label. The update strategy is shown in Algorithm 1 and α and β are experimentally set to 15 and 40, respectively.

The training of the MDN saliency detection network and pseudo-label update strategy is alternately carried to improve the performance. We regard pseudo-labels as supervision for the training saliency detection model, then use this trained model and Algorithm 1 to refine pseudo-labels, iteratively carrying the above steps until the result converges.

Algorithm 1 Pseudo label update strategy

Input: initial pseudo-labels pair Y_1 and Y_2, MDN predicted saliency pair P_1 and P_2, classification activated map R_{CAM}

Output: updated pseudo-labels pair $Y_1^{'}$ and $Y_2^{'}$

$Y_A = (Y_1 + Y_2)/2$
$P_A = (P_1 + P_2)/2$
if $MAE(Y_A, P_A) < \alpha$ **then**
    $Y_1^{'} = (P_A + Y_A)/2$
    $Y_2^{'} = Rfn((P_A + Y_A)/2)$
else if $MAE(Y_A, P_A) > \beta$ **then**
    Remove image and pseudo-labels from training data
else if $MAE(Y_A, R_{CAM}) > MAE(P_A, R_{CAM})$ **then**
    $Y_1^{'} = P_A$
    $Y_2^{'} = Rfn(P_A)$
else
    $Y_1^{'} = Y_A$
    $Y_2^{'} = Rfn(Y_A)$
end if

3.5. Testing salient object detection

The testing or inference of the saliency detection model is carried on the MDN network. Given one input image, MDN predicts two saliency results using two decoder branches and we simply adopt an average of two saliency outputs as a final saliency.

4. Experimental results

In this section, we first present the implementation details of the proposed method, then compare the performance with the state of the arts methods on four datasets and finally carry an ablation study to prove the effectiveness of our method.

4.1. Experiment setup

COLN: The weight parameters of the shared layer are initialized using the weight parameters of the pre-trained VGG model [58], while the weights of the other layers are initialized randomly. All input images are downsampled to a resolution of 256×256. To increase the training samples, we use random rotation and flipping data augmentation methods. We use small batch stochastic gradient descent (SGD) to minimize the loss function with a batch size of 64 and a momentum of 0.9. The learning rate is initialized to 0.01 and every 20 cycles decrease by a factor of 0.1. The parameter λ in Eq (3.2) is set to 2.0×10^{-5} and 2.5×10^{-5} to generate one pair of pseudo-labels. The IOU threshold parameter θ for high quality pseudo-labels selection is set to 0.6.

The training of this classification network includes two stages. In the first stage, we remove the FIN branch and only learn the parameters of the shared encoder, FCN and classification. In the second stage, we freeze the shared encoder, FCN and classification, then add the FIN branch and learn the parameters of the FIN branch. The training is carried on the NVIDIA RTX3090i hardware platform. The first stage trains 40 epochs consuming 30 hours and the second stage trains 10 epochs consuming 10 hours.

MDN: The training uses an Adam optimizer with weights decaying to 5×10^{-4} and an initial learning rate of 5×10^{-5}, divided by 10 after 15 epochs. The network was trained for a total of 24 epochs consuming four hours on the NVIDIA RTX3090i hardware platform. The backbone parameters of our network VGG-16 were initialized with the corresponding models pre-trained on the ImageNet dataset, and the rest of the models were initialized randomly. The parameter δ in Eq (3.4) is selected with two.

4.2. Datasets

Train data: Compared with image classification datasets, which have only one annotation class per object, the object detection datasets usually contain multiple objects of different classes and are more suitable for the saliency detection task. Following previous work WSS [25], the training of the COLN classification network is carried on the ImageNet object detection datasets, including 456 k objects on more than 200 object categories. Note that we only use the category labels and discard bounding box annotations.

For salient object detection model MDN, following the existing weakly supervised methods, we take DUTS-TR [25] as training set, which contains 10,553 images. Although the training set already has pixel-level annotation, we do not use pixel-level annotation.

Test data: The performance comparison was carried on a test set of four public datasets: DUTS-TE [25], ECSSD [6], PASCAL-S [59] and HKU-IS [60]. The DUTS-TE dataset contains 5019 test images, which contain important scenes for saliency detection. The ECSSD dataset contains more salient objects under complex scenes. The PASCAL-S dataset contains 850 natural images with multiple complex objects and cluttered backgrounds. The HKU-IS dataset contains 4447 images with high-quality pixel-wise annotations.

4.3. Evaluation metrics

To compare the effectiveness of these different algorithms, this work adopts two widely used evaluation metrics:

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

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

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

MAE: The difference between the saliency map S and the ground-truth G is shown below:

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

where W and H are width and height of input image, respectively.

4.4. Evaluation results

Table 2. The results quantitative evaluation on weakly supervised methods, where ↑ (↓) means the larger (smaller) value is better. The best result of category supervised methods is highlighted in bold.

Supervision	Method	DUTS-TE		PASCAL-S		ECSSD		HKU-IS	
		Max-F↑	MAE↓	Max-F↑	MAE↓	Max-F↑	MAE↓	Max-F↑	MAE↓
Fully supervised	SAMNet [61]	0.836	0.058	0.856	0.113	0.927	0.051	0.914	0.044
	PAGRN [62]	0.778	0.055	0.765	0.151	0.871	0.064	0.863	0.045
	PFAN [63]	0.764	0.060	0.754	0.137	0.875	0.046	0.871	0.042
	UCF [64]	0.663	0.112	0.787	0.140	0.844	0.070	0.818	0.062
Bounding-box	WSB [19]	0.736	0.079	-	-	0.860	0.072	0.853	0.058
Subitizing	SOS [51]	-	-	0.803	0.131	0.858	0.108	0.882	0.080
Scribble	SAE [23]	0.746	0.062	0.788	0.139	0.865	0.061	0.857	0.047
	LDS [52]	0.755	0.066	0.793	0.094	0.873	0.056	0.860	0.048
Category	WSS [25]	0.630	0.099	0.697	0.184	0.823	0.109	0.821	0.084
	ASMO [20]	0.614	0.116	0.752	0.145	0.837	0.112	0.846	0.088
	MSW [53]	0.684	0.091	0.713	0.133	0.840	0.096	0.814	0.084
	MFNet [21]	0.710	0.076	0.751	0.115	0.854	0.084	0.851	0.059
	NSAL [22]	0.729	0.074	0.753	**0.111**	0.853	0.081	0.859	0.053
	Ours	**0.749**	**0.067**	**0.785**	0.125	**0.871**	**0.063**	**0.861**	**0.049**

We compare our approach with the state of the arts weakly supervised methods, which are shown in Table 2. To make fair comparison, we use the implementation or inference results provided by the authors. Our method is based on image category supervision, compared with other image category supervision based work like WSS [25], ASMO [20], MSW [53], MFNet [21] and NSAL [22], and our method demonstrates significantly better performance on all datasets except a slightly worse MAE performance in the PASCAL-S dataset. Compared with the bounding boxes supervision based work WSB [19], we also show slightly improved performance when compared with the scribble supervision based works SAE [23] and LDS [52]. Compared with the subitizing supervised method SOS [51], we show better MAE and a slightly worse FMeasure. Compared with fully supervised methods [62–64],

our method shows worse performance, while fully supervised methods need a large number of pixel-level annotations.

A few saliency quality comparison examples are shown in Figure 7. Our method achieves promised results in both simple and complex scenes. Compared to WSS [25] and ASMO [20], which both adopt a network that relies on a single pseudo-label training, and MF-Net [21], which only relies on different refinement algorithms to obtain multiple pseudo-labels, our method can greatly reduce the noise of pseudo-labels and, thus, achieves better results.

Figure 7. A visual comparison of our approach with other state of the arts methods.

4.5. Ablation studies

To further validate the effectiveness of our method, we carry an evaluation on the DUTS-TE [25] dataset with different training strategies, which are shown in Table 3.

The result of the training saliency detection network using a single pseudo-label and two pseudo-labels is denoted as MDN-SP and MDN-MP, respectively. The MDN-MP shows significantly better performance and FMeasure increased from 0.675 to 0.718, while MAE decreased from 0.118 to 0.096. Based on MDN-MP, we further jointly learn saliency and edge detection, adopt pseudo-labels updating strategy and the results are denoted as MDN-Edge and MDN-Update, respectively. Using edge joint training and updating the pseudo-label both further achieve performance improvement.

Table 3. Evaluation results of ablation experiments on DUTS-TE dataset, where ↑ (↓) means the larger (smaller) value is better.

Metrics	MDN-SP	MDN-MP	MDN-Edge	MDN-Update
Max-F↑	0.675	0.718	0.722	0.749
MAE↓	0.118	0.096	0.077	0.067

4.6. Hyper-parameter settings

4.6.1. The weight parameter λ

To select high quality pseudo-labels for the training saliency detection model, we learn COLN twice and generate two pseudo-labels for each image, the parameter λ is selected as the optimal value $\lambda = 2.0 \times 10^{-5}$ and one approximately optimal value $\lambda_2 = 2.5 \times 10^{-5}$, respectively. Then, the consistency between two pseudo-labels is calculated to measure the quality of pseudo-labels. The optimal parameter $\lambda = 2.0 \times 10^{-5}$ is selected from the classification accuracy curve in Figure 4. We carry different selection strategy on the approximately optimal value λ_2 to select high quality pseudo-labels, then train saliency detection model MDN-MP and compare their performance against baseline MDN-SP, which is shown in Table 4.

The approximately optimal value λ_2 is selected as slightly smaller, equal and slightly larger than the optimal value λ. Note that even λ_2 is equal to λ, which means even when training COLN twice with the same parameter λ, the learned two COLN models still show little difference due to the randomness of neural network training and generates two different pseudo-labels for each image. This is useful in selecting high quality pseudo-labels, leading to limited performance improvement against MDN-SP. If the approximately optimal value λ_2 is selected as slightly smaller or larger than the optimal value λ, they both achieve more performance improvement. This phenomenon can also be explained from another view. The high quality pseudo-labels are more robust to the choice of parameters, and the slight adjustment on weight parameter λ leads to less influence on a high quality pseudo-label and more influence on a low quality pseudo-label, The consistency calculation on two pseudo-labels with a slightly different weight parameter λ is helpful to select high quality pseudo-labels and leads to performance improvement.

Table 4. The weight parameter λ analysis on DUTS-TE dataset. The best results are marked in boldface.

Network	Parameter	DUTS-TE	
		Max-F↑	MAE↓
MDN-SP		0.675	0.118
MDN-MP	$\lambda = 2.0 \times 10^{-5}$, $\lambda_2 = 1.5 \times 10^{-5}$	0.710	0.099
	$\lambda = 2.0 \times 10^{-5}$, $\lambda_2 = 2.0 \times 10^{-5}$	0.702	0.108
	$\lambda = 2.0 \times 10^{-5}$, $\lambda_2 = 2.5 \times 10^{-5}$	**0.718**	**0.096**

4.6.2. *IOU* threshold parameter θ

In Section 3.2, we compute the *IOU* of each pseudo-labels pair. The pseudo-labels with *IOU* larger than one threshold θ will be regarded as high quality pseudo-labels to train the salient object detection model. In order to analyze the influence of threshold θ, we use different threshold parameter settings to select high quality pseudo-labels and train the MDN-MP model, then evaluate the model performance on the DUTS-TE dataset. The threshold θ setting, image number of training dataset and saliency detection performance are shown in Table 5. The larger threshold parameter θ means a more strict pseudo-label selection condition, which can generate pseudo-labels with a higher reliability and achieve performance improvement. On the other hand, the excessively strict condition also decreases the training image number and diversity of training data, which leads to performance drop. The parameter θ with 0.6 shows the best performance, which is selected as the best parameter.

Table 5. The *IOU* threshold parameter θ analysis on DUTS-TE dataset. The best results are marked in boldface.

θ	Training image num	DUTS-TE	
		Max-F↑	MAE↓
0.8	6464	0.637	0.124
0.6	9209	**0.718**	**0.096**
0.4	9335	0.679	0.099
0.2	9447	0.661	0.103

4.6.3. The parameter α and β in iterative learning

In Section 3.4, to further conquer the impact of noise, we propose one pseudo-labels update strategy using parameter α and β.

In order to analyze the influence of parameter α and β, we use different parameter settings to update pseudo-labels and train final model MDN-Update, then evaluate the model performance on the DUTS-TE dataset, which is shown in Table 6. The parameter α with 15 and β with 40 shows the best performance, which is selected as the best parameter.

Table 6. The parameter α and β analysis on DUTS-TE dataset. The best results are marked in boldface.

α	β	DUTS-TE	
		Max-F↑	MAE↓
10	25	0.739	0.071
10	40	0.743	0.071
10	55	0.740	0.072
15	25	0.746	0.068
15	40	**0.749**	**0.067**
15	55	0.733	0.077
20	25	0.746	0.070
20	40	0.737	0.075
20	55	0.735	0.076

5. Conclusions

In this work, we proposed one weakly supervised salient object detection method with category annotation. To this end, we proposed COLN to roughly locate the object of an image, then generated pixel-level pseudo-labels and adopted one quality check strategy to select high quality pseudo labels, which supervised the training of MDN saliency detection networks. One pseudo-labels update strategy also was presented to iteratively optimize the pseudo-labels and saliency detection model. The proposed method outperforms other image category supervision based work. Additionally, the proposed method can be applied for other computer vision tasks such as object segmentation, object tracking and video scene understanding. However, the proposed method contains many steps and it is difficult to achieve global optimization. In the future, we will consider one end-to-end method to solve this problem. Although the weakly supervised method can decrease the cost of manual annotation, it still shows a performance gap with a fully supervised method. The trade-offs between annotation cost and accuracy should also be considered in future work.

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 conflict of interest.

References

1. R. Fan, Q. Hou, M. M. Cheng, G. Yu, R. R. Martin, S. M. Hu, Associating inter-image salient instances for weakly supervised semantic segmentation, in *Proceedings of the European Conference on Computer Vision*, (2018), 367–383. https://doi.org/10.1007/978-3-030-01240-3_23
2. N. Meeboonmak, N. Cooharojananone, Aircraft segmentation from remote sensing images using modified deeply supervised salient object detection with short connections, in *International Conference on Mathematics and Computers in Science and Engineering*, (2020), 184–187. https://doi.org/10.1109/MACISE49704.2020.00040
3. X. Yao, R. Li, J. Zhang, J. Sun, C. Zhang, Explicit boundary guided semi-push-pull contrastive learning for supervised anomaly detection, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2023), 24490–24499.
4. N. Yu, H. Li, Q. Xu, A full-flow inspection method based on machine vision to detect wafer surface defects, *Math. Biosci. Eng.*, **20** (2023), 11821–11846. https://doi.org/10.3934/mbe.2023526
5. S. Hong, T. You, S. Kwak, B. Han, Online tracking by learning discriminative saliency map with convolutional neural network, in *International Conference on Machine Learning*, (2015), 597–606. https://doi.org/10.48550/arXiv.1502.06796
6. Q. Yan, L. Xu, J. Shi, J. Jia, Hierarchical saliency detection, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2013), 1155–1162. https://doi.org/10.1109/CVPR.2013.153
7. F. Perazzi, P. Krähenbühl, Y. Pritch, A. Hornung, Saliency filters: Contrast based filtering for salient region detection, in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, (2012), 733–740. https://doi.org/10.1109/CVPR.2012.6247743
8. L. Zhang, W. Chen, W. Wang, Z. Jin, C. Zhao, Z. Cai, et al., CBGRU: A detection method of smart contract vulnerability based on a hybrid model, *Sensors*, **22** (2022), 3577. https://doi.org/10.3390/s22093577
9. L. Zhang, Y. Li, T. Jin, W. Wang, Z. Jin, C. Zhao, et al., SPCBIG-EC: a robust serial hybrid model for smart contract vulnerability detection, *Sensors*, **22** (2022), 4621. https://doi.org/10.3390/s22124621
10. L. Zhang, J. Wang, W. Wang, Z. Jin, C. Zhao, Z. Cai, et al., A novel smart contract vulnerability detection method based on information graph and ensemble learning, *Sensors*, **22** (2022), 3581, https://doi.org/10.3390/s22093581
11. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2016), 770–778. https://doi.org/10.1109/CVPR.2016.90
12. Y. Li, H. Jin, Z. Li, A weakly supervised learning-based segmentation network for dental diseases, *Math. Biosci. Eng.*, **20** (2023), 2039–2060. https://doi.org/10.3934/mbe.2023094
13. F. Chen, H. Ma, W. Zhang, SegT: Separated edge-guidance transformer network for polyp segmentation, *Math. Biosci. Eng.*, **20** (2023), 17803–17821. https://doi.org/10.3934/mbe.2023791
14. Q. Feng, X. Xu, Z. Wang, Deep learning-based small object detection: A survey, *Math. Biosci. Eng.*, **20** (2023), 6551–6590. https://doi.org/10.3934/mbe.2023282

15. C. Wu, L. Chen, A model with deep analysis on a large drug network for drug classification, *Math. Biosci. Eng.*, **20** (2023), 383–401. https://doi.org/10.3934/mbe.2023018

16. X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, M. Jagersand, BASNet: Boundary-aware salient object detection, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2019), 7479–7489. https://doi.org/10.1109/CVPR.2019.00766

17. J. X. Zhao, J. J. Liu, D. P. Fan, Y. Cao, J. Yang, M. M. Cheng, EGNet: Edge guidance network for salient object detection, in *Proceedings of the IEEE International Conference on Computer Vision*, (2019), 8779–8788. https://doi.org/10.1109/ICCV.2019.00887

18. W. Wang, S. Zhao, J. Shen, S. C. Hoi, A. Borji, Salient object detection with pyramid attention and salient edges, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2019), 1448–1457. https://doi.org/10.1109/CVPR.2019.00154

19. Y. Liu, P. Wang, Y. Cao, Z. Liang, R. W. 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

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

21. Y. Piao, J. Wang, M. Zhang, H. Lu, MFNet: Multi-filter directive network for weakly supervised salient object detection, in *Proceedings of the IEEE International Conference on Computer Vision*, (2021), 4136–4145. https://doi.org/10.1109/ICCV48922.2021.00410

22. 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

23. J. Zhang, X. Yu, A. Li, P. Song, B. Liu, Y. Dai, Weakly-supervised salient object detection via scribble annotations, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2020), 12546–12555. https://doi.org/10.1109/CVPR42600.2020.01256

24. B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2016), 2921–2929. https://doi.org/10.1109/CVPR.2016.319

25. L. Wang, H. Lu, Y. Wang, M. Feng, D. Wang, B. Yin, et al., Learning to detect salient objects with image-level supervision, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2017), 136–145. https://doi.org/10.1109/CVPR.2017.404

26. X. Zhu, C. Tang, P. Wang, H. Xu, M. Wang, J. Che, et al., Saliency detection via affinity graph learning and weighted manifold ranking, *Neurocomputing*, **312** (2018), 239–250. https://doi.org/10.1016/j.neucom.2018.05.106

27. W. Zou, N. Komodakis, HARF: Hierarchy-associated rich features for salient object detection, in *Proceedings of the IEEE International Conference on Computer Vision*, (2015), 406–414. https://doi.org/10.1109/ICCV.2015.54

28. Y. Pang, X. Zhao, L. Zhang, H. Lu, Multi-scale interactive network for salient object detection, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2020), 9413–9422. https://doi.org/10.1109/CVPR42600.2020.00943
29. X. Qin, Z. Zhang, C. Huang, M. Dehghan, O. R. Zaiane, M. Jagersand, U2-Net: Going deeper with nested U-structure for salient object detection, *Pattern Recognit.*, **106** (2020), 107404. https://doi.org/10.1016/j.patcog.2020.107404
30. X. Zhao, Y. Pang, L. Zhang, H. Lu, L. Zhang, Suppress and balance: A simple gated network for salient object detection, in *Proceedings of the European Conference on Computer Vision*, (2020), 35–51. https://doi.org/10.1007/978-3-030-58536-5_3
31. L. Tang, B. Li, Y. Zhong, S. Ding, M. Song, Disentangled high quality salient object detection, in *Proceedings of the IEEE International Conference on Computer Vision*, (2021), 3580–3590. https://doi.org/10.1109/ICCV48922.2021.00356
32. M. Ma, C. Xia, J. Li, Pyramidal feature shrinking for salient object detection, in *Proceedings of the AAAI Conference on Artificial Intelligence*, **35** (2021), 2311–2318. https://doi.org/10.1609/aaai.v35i3.16331
33. Y. Song, H. Tang, M. Zhao, N. Sebe, W. Wang, Quasi-equilibrium feature pyramid network for salient object detection, *IEEE Trans. Image Process.*, **31** (2022), 7144–7153. https://doi.org/10.1109/TIP.2022.322005
34. M. Zhuge, D. P. Fan, N. Liu, D. Zhang, D. Xu, L. Shao, Salient object detection via integrity learning, *IEEE Trans. Pattern Anal. Mach. Intell.*, **45** (2022), 3738–3752. https://doi.org/10.1109/TPAMI.2022.3179526
35. Z. Wu, S. Li, C. Chen, H. Qin, A. Hao, Salient object detection via dynamic scale routing, *IEEE Trans. Image Process.*, **31** (2022), 6649–6663. https://doi.org/10.1109/TIP.2022.3214332
36. Y. H. Wu, Y. Liu, L. Zhang, M. M. Cheng, B. Ren, EDN: Salient object detection via extremely-downsampled network, *IEEE Trans. Image Process.*, **31** (2022), 3125–3136. https://doi.org/10.1109/TIP.2022.3164550
37. M. Ma, C. Xia, C. Xie, X. Chen, J. Li, Boosting broader receptive fields for salient object detection, *IEEE Trans. Image Process.*, **32** (2023), 1026–1038. https://doi.org/10.1109/TIP.2022.3232209
38. R. Bi, C. Ji, Z. Yang, M. Qiao, P. Lv, H. Wang, Residual based attention-unet combing DAC and RMP modules for automatic liver tumor segmentation in CT, *Math. Biosci. Eng.*, **19** (2022), 4703–4718. https://doi.org/10.3934/mbe.2022219
39. H. Zhu, X. He, M. Wang, M. Zhang, L. Qing, Medical visual question answering via corresponding feature fusion combined with semantic attention, *Math. Biosci. Eng.*, **19** (2022), 10192–10212. https://doi.org/10.3934/mbe.2022478
40. C. Jin, J. Huang, T. Wei, Y. Chen, Neural architecture search based on dual attention mechanism for image classification, *Math. Biosci. Eng.*, **20** (2023), 2691–2715. https://doi.org/10.3934/mbe.2023126
41. M. Chen, S. Yi, M. Yang, Z. Yang, X. Zhang, Unet segmentation network of COVID-19 CT images with multi-scale attention, *Math. Biosci. Eng.*, **20** (2023), 16762–16785. https://doi.org/10.3934/mbe.2023747

42. N. Liu, N. Zhang, K. Wan, L. Shao, J. Han, Visual saliency transformer, in *Proceedings of the IEEE International Conference on Computer Vision*, (2021), 4722–4732. https://doi.org/10.1109/ICCV48922.2021.00468

43. Z. Wang, P. Wang, Y. Han, X. Zhang, M. Sun, Q. Tian, Curiosity-driven salient object detection with fragment attention, *IEEE Trans. Image Process.*, **31** (2022), 5989–6001. https://doi.org/10.1109/TIP.2022.3203605

44. C. Xie, C. Xia, M. Ma, Z. Zhao, X. Chen, J. Li, Pyramid grafting network for one-stage high resolution saliency detection, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2022), 11717–11726. https://doi.org/10.1109/CVPR52688.2022.01142

45. D. P. Fan, J. Zhang, G. Xu, M. M. Cheng, L. Shao, Salient objects in clutter, *IEEE Trans. Pattern Anal. Mach. Intell.*, **45** (2022), 2344–2366. https://doi.org/10.1109/TPAMI.2022.3166451

46. M. M. Cheng, S. H. Gao, A. Borji, Y. Q. Tan, Z. Lin, M. Wang, A highly efficient model to study the semantics of salient object detection, *IEEE Trans. Pattern Anal. Mach. Intell.*, **44** (2021), 8006–8021. https://doi.org/110.1109/TPAMI.2021.3107956

47. X. Tian, K. Xu, X. Yang, L. Du, B. Yin, R. W. Lau, Bi-directional object-context prioritization learning for saliency ranking, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2022), 5882–5891.

48. X. Tian, X. Yang, B. Yin, R. W. Lau, Weakly-supervised salient instance detection, preprint, arXiv:2009.13898.

49. X. Tian, K. Xu, X. Yang, B. Yin, R. W. Lau, Learning to detect instance-level salient objects using complementary image labels, *Int. J. Comput. Vision*, **130** (2022), 729–746. https://doi.org/10.1007/s11263-021-01553-w

50. Z. Liang, P. Wang, K. Xu, P. Zhang, R. W. Lau, Weakly-supervised salient object detection on light fields, *IEEE Trans. Image Process.*, **31** (2022), 6295–6305. https://doi.org/10.1109/TIP.2022.3207605

51. X. Zheng, X. Tan, J. Zhou, L. Ma, R. W. H. Lau, Weakly-supervised saliency detection via salient object subitizing, *IEEE Trans. Circuits Syst. Video Technol.*, **31** (2021), 4370–4380. https://doi.org/10.1109/TCSVT.2021.3049408

52. X. Liu, J. Guo, S. Zheng, Weakly-supervised salient object detection with label decoupling siamese network, in *Proceedings of the 8th International Conference on Computing and Artificial Intelligence*, (2022), 412–418. https://doi.org/10.1145/3532213.3532275

53. Y. Zeng, Y. Zhuge, H. Lu, L. Zhang, M. Qian, Y. Yu, Multi-source weak supervision for saliency detection, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2019), 6074–6083. https://doi.org/10.1109/CVPR.2019.00623

54. H. Zhang, Y. Zeng, H. Lu, L. Zhang, J. Li, J. Qi, Learning to detect salient object with multi-source weak supervision, *IEEE Trans. Pattern Anal. Mach. Intell.*, **44** (2021), 3577–3589. https://doi.org/10.1109/TPAMI.2021.3059783

55. C. Rother, GrabCut: interactive foreground extraction using iterated graph cuts, *ACM Trans. Graphics*, **23** (2004), 309–314. https://doi.org/10.1145/1015706.1015720

56. Y. Boykov, M. Jolly, Interactive graph cuts for optimal boundary and region segmentation of objects in n-d images, in *Proceedings of the IEEE International Conference on Computer Vision*, (2001), 105–112. https://doi.org/10.1109/ICCV.2001.937505

57. J. J. Liu, Q. Hou, M. M. Cheng, J. Feng, J. Jiang, A simple pooling-based design for real-time salient object detection, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2019), 3917–3926. https://doi.org/10.1109/CVPR.2019.00404

58. K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, preprint, arXiv:1409.1556.

59. Y. Li, X. Hou, C. Koch, J. M. Rehg, A. L. Yuille, The secrets of salient object segmentation, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2014), 280–287. https://doi.org/10.1109/CVPR.2014.43

60. G. Li, Y. Yu, Visual saliency based on multiscale deep features, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2015), 5455–5463. https://doi.org/10.1109/CVPR.2015.7299184

61. Y. Liu, X. Y. Zhang, J. W. Bian, L. Zhang, M. M. Cheng, SAMNet: Stereoscopically attentive multi-scale network for lightweight salient object detection, *IEEE Trans. Image Process.*, **30** (2021), 3804–3814. https://doi.org/10.1109/TIP.2021.3065239

62. X. Zhang, T. Wang, J. Qi, H. Lu, G. Wang, Progressive attention guided recurrent network for salient object detection, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2018), 714–722.

63. T. Zhao, X. Wu, Pyramid feature attention network for saliency detection, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2019), 3085–3094.

64. P. Zhang, D. Wang, H. Lu, H. Wang, B. Yin, Learning uncertain convolutional features for accurate saliency detection, in *Proceedings of the IEEE International Conference on Computer Vision*, (2017), 212–221.

AIMS Press