Complementary characteristics fusion network for weakly supervised salient object detection

Published in:
Image and Vision Computing

Publication Status:
Published (in print/issue): 31/10/2022

DOI:
10.1016/j.imavis.2022.104536

Document Version
Author Accepted version

Noname manuscript No.
(will be inserted by the editor)

Complementary Characteristics Fusion Network for Weakly Supervised Salient Object Detection

Yan Liu · Yunzhou Zhang* · Zhenyu Wang · Fei Yang · Cao Qin · Feng Qiu · Sonya Coleman · Dermot Kerr

Received: date / Accepted: date

Abstract Salient object detection is a challenging and fundamental research in computer vision and image processing. Although the fully convolutional network has made a great progress in the saliency detection task, most existing methods mainly rely on dense ground truth as labels for training, which takes extensive effort and is time-consuming. This paper proposes a novel and effective scribble-based weakly supervised approach named complementary characteristics fusion network (CCFNet), which learns from easily accessible scribbles such as centerlines instead of fully pixel-wise ground truth. To be more specific, in order to deal with the fact that scribbles are always located inside the objects with lacking annotations close to the semantic boundaries, an edge fusion module is presented to equip our model with the power of aggregating edge information, which would be beneficial to generate saliency maps with more useful information. Alternatively, since scribbles are too sparse to provide enough supervision for the network, we design feature correlation modules based on low-level, high-level global and edge information, which will complement each other to obtain relatively complete salient regions using features of different ways. To further improve the results of saliency maps in foreground and background, a self-supervised saliency detection loss is designed to ensure the network with stronger generalization ability. Extensive experiments using five benchmark datasets demonstrate that our proposed approach performs favorably against the state-of-the-art weakly supervised algorithms, and even surpasses the performance of those fully supervised.

Keywords Salient object detection, Weakly supervised learning, Complementary characteristics fusion network, Self-supervised saliency detection loss

* Corresponding author.
Y. Liu, and Z. Wang are with Faculty of Robot Science and Engineering, Northeastern University, Shenyang, China.
Y. Zhang, C. Qin and F. Qiu are with College of Information Science and Engineering, Northeastern University, Shenyang, China (E-mail: zhangyunzhou@mail.neu.edu.cn(✉)).
F. Yang is with Zhejiang Lab new campus, Zhongtai, Yuhang, Hangzhou, China.
S. Coleman and D. Kerr are with School of Computing, Engineering and Intelligent Systems, Ulster University, N. Ireland, UK.

1 Introduction

The objective of salient object detection (SOD) is to locate and segment the most dominant objects in a given image. It plays an important role in a variety of computer vision and image processing related fields, such as image manipulation [5,10], robot navigation [6], semantic segmentation [39] and object tracking [53,11].

Following the previous studies, fully deep learning methods have been developed, which broke the limits of traditional handcrafted features since their capability of extracting features at various scales. However, these deep learning based methods usually suffer from a key problem, that they strongly depend on a large volume of accurately labeled data with full pixel-wise annotations for training. It takes extensive effort and time to collect. Therefore, this paper concentrates on designing weakly supervised salient object detection methods based on the sparse labels.

In order to address a trade-off between label efficiency and model performance, some researchers attempted to develop a framework to learn saliency maps from the sparse label [31,45,49,48,21], but there still remains challenges. Image-level category labels are used in [34], which requires large scale datasets with image-level labels. A related work [21] utilized bounding box labels

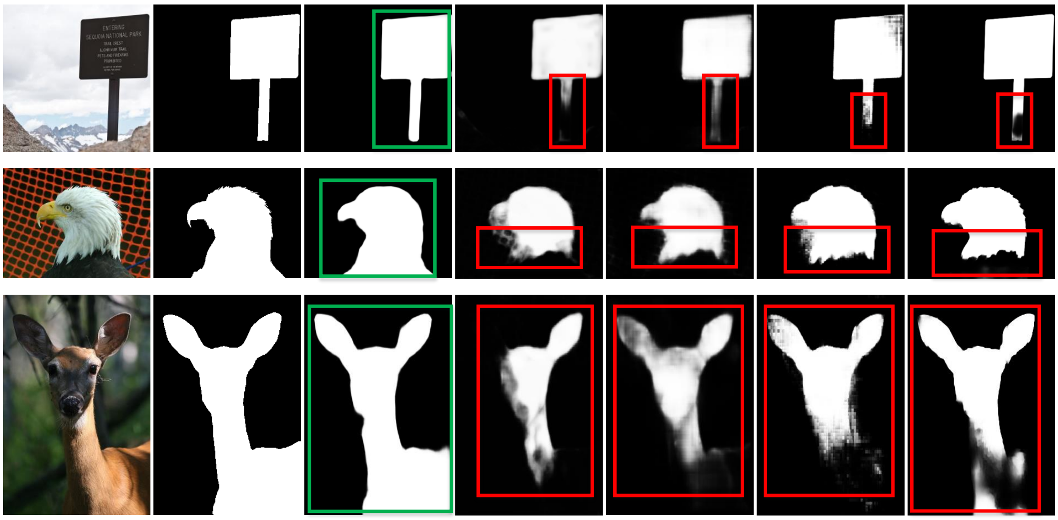

Fig. 1 Sample results of our method compared with other unsupervised and weakly supervised methods.

as supervision, which first produced the initial pseudo ground truth saliency maps by unsupervised learning, then adopted post-processing to obtain the final dense predictions. As shown in Fig. 1, WSS (image-level category labels) and SBBs (bounding box labels) only can detect part of salient regions with wrong errors. For example, given an image with an eagle in Fig. 1, the forementioned approaches are only able to segment the head of the eagle (white) whereas the body is also a salient object. Moreover, scribble annotations [48, 27] are becoming more and more popular in computer vision, which belongs to a middle ground between image-level supervision and box-level supervision. The key problem for saliency detection based on scribble annotations lies in two aspects. The first one is that the scribbles are always located inside the objects with lacking annotations close to the semantic boundaries, and thus usually generate imprecise saliency maps on boundaries. The second one is the scribbles are too sparse to provide enough supervision information for the network, which can't make confident predictions.

With respect to the first issue, as illustrated in [41, 55], edge information has been widely used and has made a great progress in fully supervised saliency detection, weakly supervised SOD models rarely have such ideas. Therefore, edge fusion module (EFM) is employed to capture edge information from local and global views, instead of the simple backbone features, which can effectively improve edge performance of saliency maps. To alleviate the second issue, we propose a feature correlation module (FCM) to capture rich information. Note that different level features have different functions, such as low-level features have rich details and high-level features have rich semantics. Our FCM achieves complement each other of different input that has a large potential to exploit the relationship from different views. The work in [48] which is different from our approach, only use concatenation operation to fuse different features. Our FCM correlates low-level features, high-level global features and edge features at different stages, which is conductive to enhancing the saliency maps. Furthermore, structural information is also crucial for scribble supervised SOD except for context information. Inspired from [1], we design a self-supervised saliency detection loss to learn structural information, which ensures the network with stronger generalization ability and distinguishes the foreground and background better. As shown in Fig. 1, benefiting from the above, our proposed approach is able to detect more accurate edge information with some challenging environments compared with other methods, such as low contrast scenarios (the background behind the deer) or complex scene understanding (details of the sign).

Based on the above consideration, we propose a complementary characteristics fusion network (CCFNet) for weakly supervised salient object detection. We design edge fusion module to learn salient edge information, which can better understand edge information. In order to exploit complete salient regions with different level features, this paper proposes feature correlation modules for saliency detection. Meantime, the output of global context guiding operation is fed into feature correlation module as input, which could address the high-level features gradually diluted as the top-down pathways. To boost the performance of our proposed model, a self-supervised saliency detection loss is presented as well to distinguish foreground and background. Finally, to demonstrate the performance of our proposed method, we conduct experiment results on five well-known datasets. Some ablation studies are reported as well to evaluate the effect of each module. From the above, our main contributions can be summarized as follows:

- We develop a novel complementary characteristics fusion network (CCFNet) based on scribble annotations for salient object detection, without resorting to laborious pixel labeling.
- We propose an edge fusion module to equip our model with the power of aggregating edge information. In addition, a feature correlation module is employed to make full use of the complementarities different features to improve saliency detection accuracy.
- We introduce a self-supervised saliency detection loss, which encourages our network to learn structural information and guides the network paying high attention to saliency objects.
- Experimental results demonstrate that the proposed approach achieves comparable performance on five common datasets compared with other state-of-the-art methods, where it even performs comparably to some of the fully supervised methods.

2 Related Work

Fully supervised salient object detection Traditional SOD approaches mainly depend on some hand-crafted features [2,4,14,46] to directly detect salient objects in each image while lacking in high-level semantic information, especially in the complex environments. Compared with early researches on SOD, deep learning based methods [17,28,56,12,30,35,36,37,23,3,54,26,33,22] have become popular because of their accurate performance. On the one hand, a variety of effective fully convolutional network based (FCN-based) architectures [23,13,36,37] have been proposed to enhance the generation of saliency maps in literature. For example, Hou et al. [13] utilized short connections for multi-scale feature fusion from different layers in FCN to address the scale-space problem. In [36], Wang et al. employed fixation prediction to segment salient objects in an attentive saliency network (ASNet), demonstrating that ASNet achieves more accurate results due to the computed fixation map. The F^3Net was introduced in [37], to solve the problem generated by the different receptive fields of different convolutional layers. On the other hand, edge information has been attracted attention to assist the performance of saliency prediction [41,55,19,38]. Zhao et al. [55] designed an edge guidance network for salient object detection with binary cross-entropy. Liu et al. [19] adopt other edge datasets as ground truth for joint training. [55,41] used edge ground-truth as auxiliary supervision, it proves that is helpful for saliency maps, especially object boundaries. Moreover, a number of related works [3,54,56] leveraged the attention mechanism to learn more distinctive features, others [12,37] introduced multi-level features to boost the performance of saliency maps. Although these methods achieve highly-accurate results, deep models require a large number of fully annotated images when trained on datasets, which is a labor-intensive and costly process.

Weakly supervised salient object detection To reduce the time and the cost of labeling, weakly-supervised learning utilize weak labels for the saliency detection task, such as noisy labels [25,47,49], bounding boxes [31], scribble annotations [44,48] and image-level labels [15,34], which have received a lot of attention from researchers. Currently, Wang et al. [34] adopted foreground inference network for object saliency prediction with image-level labels, which is the first application of image-level labels to SOD. Li et al. [15] subsequently introduced a multi-task fully convolutional network (Multi-FCN) to generate saliency maps using image-level weak supervision. Piao et al [29] built a saliency network and multiple directive filters to enhance the performance of SOD, which is a multiple-pseudo-label framework. Furthermore, S-DUTS was proposed in [48] first on saliency detection, which is based on sparse labels and typically takes 1∼2 seconds to label each image. They also designed a network fusing edge detection approach and a gated structure-aware loss function to maintain the accuracy of the salient prediction. Yu et al. [44] introduced a one-round end-to-end training approach using scribbles for weakly-supervised saliency detection. Unlike these methods, we cooperate the characteristics and complementarity of different features to reduce the gap between fully supervised learning and weakly-supervised learning.

Unsupervised salient object detection Early methods have been proposed for predicting the saliency map, which mainly used some handcrafted features, contrast, different priors and so on [10,14,4]. A related work [47] proposed a deep learning framework from unsupervised methods with heuristics to produce saliency maps. Li et al. [18] developed a contour-to-saliency network based on the well-trained contour detection network. Subsequently, Nguyen et al. [25] presented a two-stage network for unsupervised saliency detection to improve prediction quality, which was updated through noisy labels generated. In conclusion, ground truth is not required for these methods. Unsupervised learning on salient object detection has been made a great and

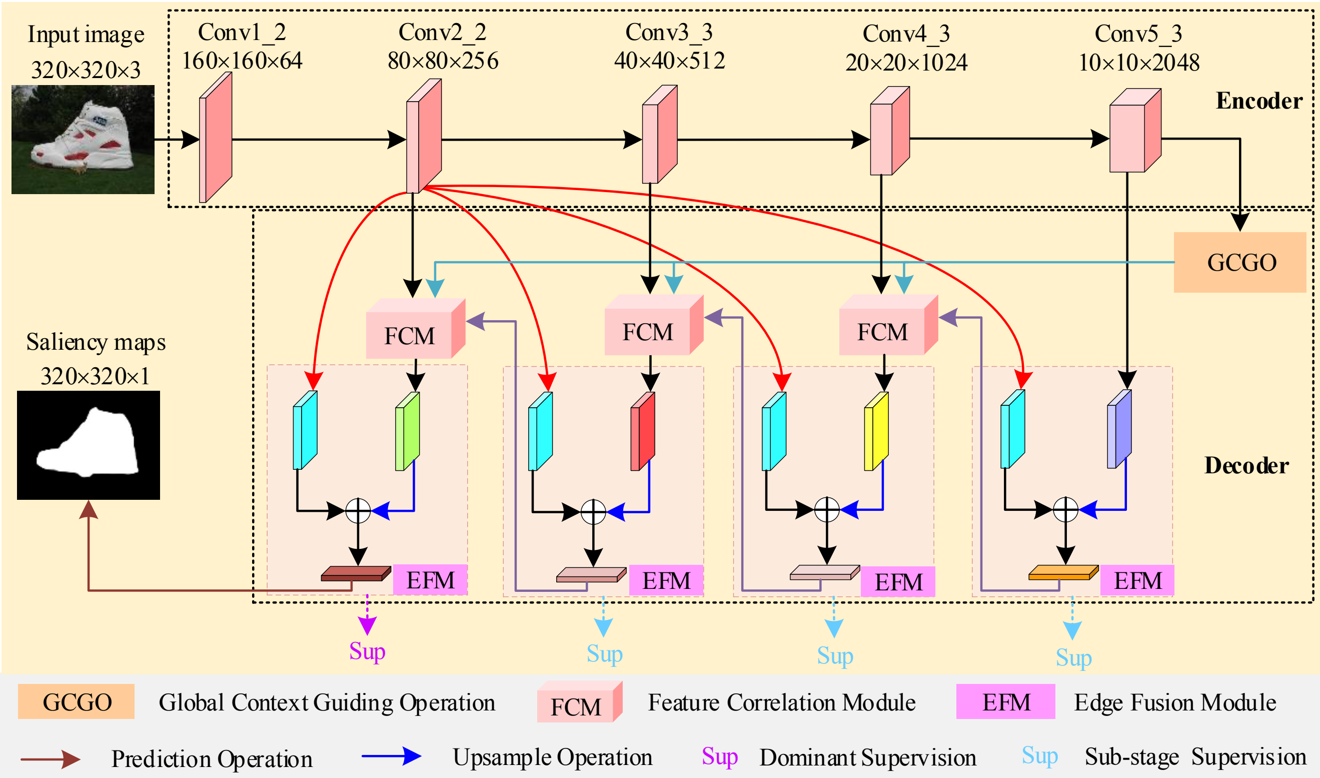

Fig. 2 Illustration of our proposed CCFNet architecture.

significant process, but the accuracy is limited due to the gap between fully supervised learning and unsupervised learning.

3 Methodology

In this section, we first introduce the proposed complementary characteristics fusion network (CCFNet) for weakly supervised salient object detection. Then the details of the global context guiding operation (GCGO), the edge fusion module (EFM) and the feature correlation module (FCM) are described. The network supervision strategy is presented at the end of this section.

3.1 Overall pipeline

The overall architecture of CCFNet is illustrated in Fig. 2. Our model is designed based on FCN architecture and chooses ResNet-50 as the backbone, which consists of five convolutional blocks for feature extracting. Given an input image with size $H \times W$, the encoder will generate different level features, denoted as $\{f_i|i=1,\cdots,5\}$ with resolutions $[\frac{H}{2^i}, \frac{W}{2^i}]$. Since the 1st level feature f_1 would increase computation cost and have a lot of noises, which yields limited performance improvements, we choose features from $\{f_i|i = 2,\cdots,5\}$ for later operations. Specifically, to alleviate the problem of U-shape networks as top-down ways gradually diluted, f_5 is fed into GCGO to obtain $\{g_i|i = 1,\cdots,3\}$, which can guarantee global semantics delivered. Since low-level features have more details such as boundaries, which are useful and indispensable for generating accurate saliency maps, we leverage the f_2 to extract the boundaries. In contrast, high-level features have more semantics but lacking details. Taking account of these considerations, we aim to explicitly notice the salient edges where salient objects are. Hence we model EFM to strengthen edge information, denoted as $\{e_i|i = 1,\cdots,4\}$. Besides, in view of different types of features delivering different information, to this end, we design FCM in this paper. FCM is performed to refining low-level features $\{f_i|i = 2,\cdots,4\}$, global high-level features $\{g_i|i = 1,\cdots,3\}$ and edge features $\{e_i|i = 1,\cdots,4\}$. This way enables the network to understand scenarios from different views, which will generate the discriminative features. It may limit the capability of the network due to only choosing scribbles to train our network. To address this limitation, we also propose a self-supervised saliency detection loss for joint training to enrich structural information. More details of CCFNet are described as follows.

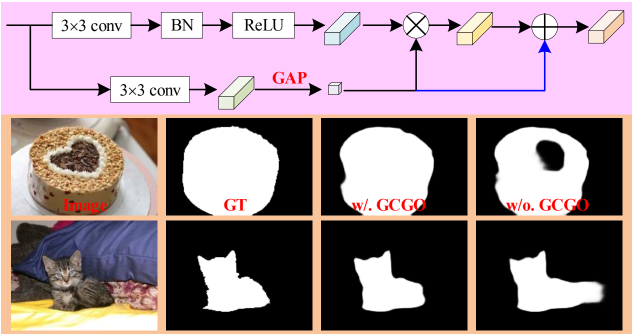

Fig. 3 Illustration of our proposed global context guiding operation (GCGO).

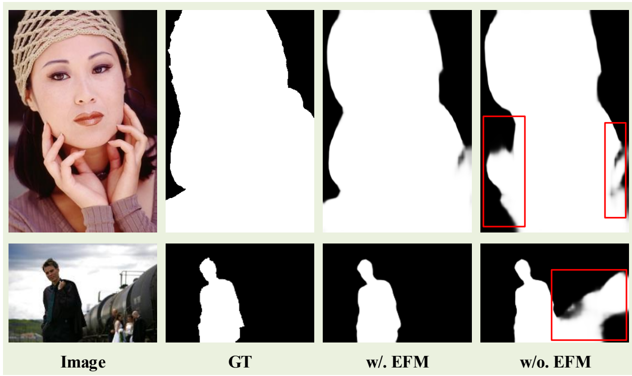

Fig. 4 Visual results by applying EFM and without EFM.

3.2 Global context guiding operation

Regarding the U-shape architecture exists an issue that the high-level features will be gradually diluted as the top-down pathways. Therefore, we propose a global context guiding operation to strengthen high-level information and obtain global information, as shown in Fig. 3. Specifically, we apply a combination of 3×3 convolutional $\rightarrow$ batch normalization $\rightarrow$ ReLU operation for input feature f_5. After that, a global average pooling (GAP) layer is embedded on these features, denoted as f_g, which can capture a more robust spatial translations of the input and the strongest global context. The refined feature f_g is denoted as follows:

$$f_g = GAP(\sigma(\phi(Conv(f_5, \theta)))), \tag{1}$$

where each of $Conv(\cdot, \theta)$, denotes the convolution with parameter θ, $\phi(\cdot)$, $\sigma(\cdot)$ and $GAP(\cdot)$ denotes the batch normalization, Relu and global average pooling, respectively. Meantime, we apply 3×3 convolutional operation to input features to squeeze the input feature f_5 and adopt a upsample operation, which retain useful information. Finally, we generate the mask $\mathbf{W}$ and bias $\mathbf{b}$ for multiplication and addition operation. The whole process is formally formulated as follows.

$$g_1 = \sigma(\mathbf{W} * f_g + \mathbf{b}), \tag{2}$$

where $*$ is element-wise multiplication and g_1 is the final output. From Fig. 3, we can clearly see that with GCGO strategy achieves better performance than without it. Detailed quantitative studies of GCGO can be found in Section 4.

3.3 Edge fusion module

Ideally, a good weakly supervised salient object detection algorithm should have the ability to capture accurate edge information. In other words, salient edge result is able to help salient object detection tasks in both segmentation and localization. To this end, we propose a series of edge fusion modules (EFM) to model the salient edge information. As stated before, the f_2 retains edge information, even so, it is still local information and not enough. We take account of high-level semantics, which are essential and necessary for obtaining salient edge information as well.

To be more specific, taking the first EFM as an example, we take a 3×3 convolutional layer after extracting the edge feature from f_2. In order to increase the reliability of salient edge information, we fuse high-level semantic information from f_5. We add a convolutional operation with kernel size 3×3 after fusing both features, it is able to effectively reduce the aliasing effect of upsampling. For the other EFMs, our goal is the high level cue mined is applied over the corresponding feature, which is further propagated to the next EFM for edge generation. That is to say, the feature maps from the corresponding feature correlation module are replaced of the backbone feature from f_5. The whole process is formally formulated as follows.

$$h_i = \begin{cases} Conv(f_2, \theta) + Up(f_5), & if\ i = 1 \\ Conv(f_2, \theta) + Up(\Psi_i), & if\ i = 2, 3, 4 \end{cases} \tag{3}$$

where Ψ_i is the output of feature correlation module. Hence the final output of EFM can be described as follows.

$$e_i = \theta(\phi(Conv(h_i, \theta)))\ i = 1, 2, 3, 4 \tag{4}$$

where e_i is the final output. To verify the effectiveness of our designed EFM, we visualize the saliency maps by applying EFM or not in Fig 4, it is clearly observed that our model with EFM has high quality in edge.

3.4 Feature correlation module

Given three pathway features: low-level features, global high-level features and edge features in CCFNet, which

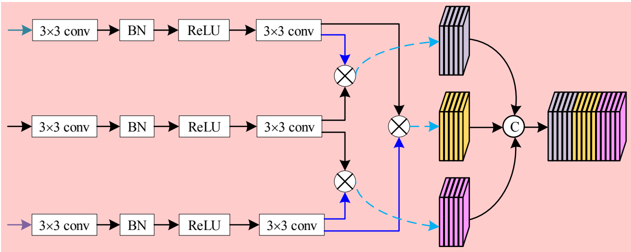

Fig. 5 Illustration of our proposed feature correlation module (FCM).

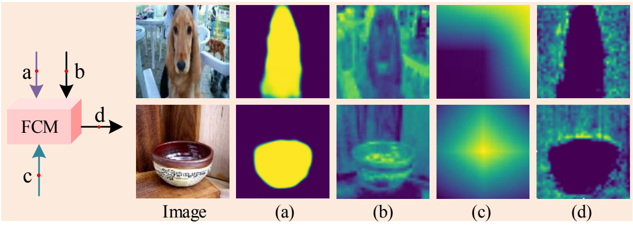

Fig. 6 Visualization of the feature maps around FCM. (a) Results of applying EFM. (b) Results of applying backbone. (c) Results of applying GCGO. (d) The output results of FCM.

can better preserve details, global semantics and edge information, respectively. However, there exists a issue that a single feature provides locally limited information. To solve this deficiency, considering that these features are complementary to each other, it is essential to form an effective decoder to strengthen the quality of saliency maps.

To this end, we define the feature correlation module (FCM) to get rich features from different pathway in this section, which is able to produce saliency maps with accurate segmentation. The details of FCM is illustrated in Fig. 5. Formally, we first conduct a series of operations: $Conv(3\times 3,\ 256) \rightarrow BN \rightarrow ReLU \rightarrow Conv(3\times 3,\ 256)$, denoted as F_i, G_i and E_i respectively. Inspired by the attention mechanism [3], we evaluate the interaction between the any embedding features that is used to generate a global feature by aggregating every local feature. Therefore, we adopt upsample to feature G_i (cyan line as input) firstly so that it has the same size as feature F_i (black line as input), then the mutual influence of feature G_i and feature F_i is achieved by element-wise multiplication, that is $Up(G_i) * F_i$. This way is used to capture more discriminative characterastics representation from global context and details. On the one hand, it is able to obtain features from details and boundaries, on the other hand, it can gain features in global dimension and in edge. Note that these features complement each other to form feature correlation module with more discerning capabilities, which are critical for salienct detection. The whole process is defined as follows.

$$\begin{cases} \Upsilon_i = Cat(Up(G_i) * F_i, \\ \qquad Up(E_i) * F_i,\ Up(E_i * G_i)), \quad if\ i = 1,2,3 \\ \Psi_i = Conv(\Upsilon_i, \theta), \end{cases} \tag{5}$$

where $Cat(\cdot,\cdot)$ denotes concatenate operation and Ψ_i denoted the output of ith FCM, respectively. Furthermore, to verify the rationality of our proposed FCM, we visualize the feature maps near the FCM in Fig. 6, which can see that FCM are helpful and combining them together are able to remedy for the deficiency of each branch feature. Detailed quantitative studies of FCM can be found in Section 4.

3.5 Self-supervised saliency detection loss

In view of the fact that we only choose scribble annotations to train our network, which contains a large number of unlabeled pixels and thus may limit the capability of the network. The proposed modules focus on obtaining context information, whereas structural information also plays an important role in scribble supervised saliency object detection. Partial cross-entropy (PCE) loss [32] is widely used to weakly supervised learning. However, it only calculates binary cross-entropy loss between the scribbles and the predicted map while not comprehensive for saliency detection. Based on this consideration, to encourage better saliency maps with more structural information from the network, we propose a novel self-supervised saliency detection (SSD) loss to help the network better distinguish foreground and background.

As shown in Fig. 7, it can be clearly seen that the results of the first two phases (d & e) are not as good as that of the third (f). Considering that the later EFM maintains more information than the former, and the results of the last EFM is the prediction result of the whole network. To learn more structural information and guide the network paying high attention to saliency objects, we produce pseudo ground truth masks from the penult EFM by considering confidence $>$ %60 (that is MAE scores < 0.6), which generate saliency maps that are closer near to ground truth, rather than scribbles. Note that the pixels with low confidence are ignored by the loss function. To this end, we design a gate function to judge if it satisfy the needs of the above. The details are described as follows.

$$g(x,y) = \begin{cases} 1, & if\ PSE_{MAE}(x,y) < 0.6 \\ 0, & otherwise \end{cases} \tag{6}$$

where $PSE_{MAE}(x,y)$ is the MAE scores of the results from the penult EFM. Motivated by semantic segmentation method [1], we use pixels cross-entropy loss, but

Table 1 Main characteristics of the datasets used in the experiments.

Name	Stage	Size	Description
S-DUTS [48]	Train	10553	Relabel salient object detection dataset DUTS-TR with scribbles.
ECSSD [42]	Test	1000	A dataset includes semantically meaningful and complex structures.
DUT-OMRON [43]	Test	5168	A dataset with high quality and challenging images has one or more salient objects with complex background scenes.
PASCAL-S [7]	Test	850	A dataset selected from the PASCAL VOC 2010 segmentation, which contains 20 object categories and complex scenes.
HKU-IS [16]	Test	4447	Contain multiple salient objects with overlapping objects touching the image boundary or with low color contrast.
DUTS-TE [34]	Test	5019	It selected from the largest salient object detection benchmark dataset DUTS, which contains complex scens in different scales.

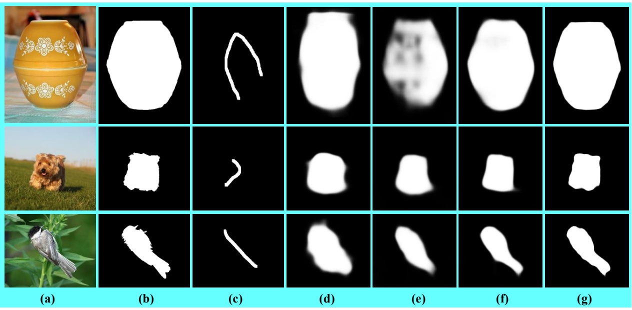

Fig. 7 Intermediate results at training time. (a) Input image. (b) Per-pixel wise ground truth. (c) Scribble annotations. (d) Results of applying first EFM. (e) Results of applying second EFM. (f) Results of applying third EFM (pseudo ground truth). (g) Ours.

the loss for saliency objects are normalized according to the number of corresponding pixels contained in the pseudo ground truth. Hence, SSD loss can be described as follows.

$$L_{ssd} = g(x,y)\ L_{bce}, \tag{7}$$

where

$$L_{bce} = -\sum_{(x,y)} [p(x,y)\log(q(x,y)) + (1-p(x,y))\log(1-q(x,y))], \tag{8}$$

where $p(x,y)$ and $q(x,y)$ denote the pseudo ground truth masks and the predicted saliency maps, respectively.

3.6 Objective Function

Given an input image, we utilize the loss of each sub-stage and the dominant loss to train our model. First, the loss of sub-stage is defined as follows.

$$L_{sub} = L_{pce} + L_{lsc}, \tag{9}$$

where

$$L_{pce} = \sum_{i\in S} -s_i \log \hat{s}_i + (1-s_i)\log(1-\hat{s}_i), \tag{10}$$

$$L_{lsc} = \sum_{i} \sum_{j\in G_i} F(i,j)D(i,j), \tag{11}$$

Here, Eq. (10) is partial cross-entropy (PCE) loss, which is widely used to weakly supervised learning, where s is the scribble annotations, $\hat{s}$ is the predicted values and S is the labeled pixel set. However, due to PCE loss only calculating binary cross-entropy loss between the scribbles and the predicted map, it is not comprehensive for saliency detection. In order to further use scribble annotations, local saliency coherence (LSC) loss Eq. (11) is adopted in previous work [44], where $F(i,j)$ is Gaussian kernels and $D(i,j)$ is $L1$ distance. Second, self-supervised saliency detection will joint with PCE loss and LSC loss to supervise in this paper. The dominant loss can be described as follows.

$$L_{dom} = L_{pce} + L_{lsc} + L_{ssd}. \tag{12}$$

Hence, the total loss in the whole network can be expressed as follows.

$$L = L_{dom} + \gamma_i \sum_{i=1}^{3} L^i_{sub}, \tag{13}$$

where γ_i is the a coefficient to balance the dominant loss and the different sub-stage loss. Because different sub-stage provides various extend of information, we set $\gamma_1 = 0.8, \gamma_2 = 0.6, \gamma_3 = 0.4$ in this paper.

4 Experiments

4.1 Implementation details

The proposed approach was implemented on the Pytorch platform using a RTX3090 GPU. The batch size is set as 16. The whole network is optimized by stochastic gradient descent (SGD), where the weight decay is set to 5e-4, the momentum is set to 0.9 and the initial learning rate 1e-5. We resize each image to 320×320 and then feed into the network to obtain saliency maps. Additionally, the characteristics of each dataset are summarized in Table 1.

4.2 Evaluation metrics

For the salient object detection task, six popular evaluation metrics are used to evaluate the effectiveness of our CCFNet including precision-recall curve (PR curve), F-measure curve, F-measure score(F_β), mean absolute error (MAE), E-measure score (E_ϕ) and S-measure score (S_α).

PR curve can be determined by generated pairs of precision and recall values. Precision and recall are computed as:

$$Precision = \frac{TP}{TP+FP}, \quad Recall = \frac{TP}{TP+FN}, \tag{14}$$

where TP, FP and FN denote true-positive, false-positive and false-negative, respectively.

F-measure score(F_β) is an overall performance measurement, which is calculated by the weighted harmonic mean of precision and recall:

$$F_\beta = \frac{(1+\beta^2) \times Precision \times Recall}{\beta^2 \times Precision + Recall}, \tag{15}$$

where $Precision$ and $Recall$ are given by thresholding the predicted saliency map, and β^2 is set to 0.3 in accordance with [3]. Then the obtained pairs (threshold, F_β) is employed to plot the F-measure curve.

MAE reflects the average pixel-wise absolute difference between the saliency map $S(x, y)$ and ground-truth maps $G(x, y)$:

$$MAE = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} |S(x, y) - G(x, y)|, \tag{16}$$

where W and H represent width and height of the saliency maps respectively. This is an appropriate metric for evaluating the applicability of a saliency module in a task such as image segmentation.

Enhanced-alignment measure E_ϕ [9] is applied to evaluate both local and global similarity between the predicted map and the ground-truth:

$$E_\phi = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} F_\phi(x, y), \tag{17}$$

F_ϕ denotes the enhanced alignment matrix.

Structure measure S_α [8] is utilized as the structure similarity of the predicted non-binary saliency map and the ground-truth, which is defined as follows:

$$S_\alpha = (1-\alpha)S_r + \alpha S_o, \tag{18}$$

where S_r and S_o denote region-aware and object-aware structural similarity respectively, and α is typically set to 0.5.

4.3 Comparison with the State-of-the-Art Methods

We compare our model with state-of-the-art nineteen methods, including eleven fully supervised methods (Amulet [51], UCF [52], NLDF [24], RAS [3], PAGR [54], BMPM [50], DSS [12], EGNet [55], CPD [40], MINet [28] and VST [20]), two unsupervised methods (SVF [47] and C2S [18]), six weakly supervised methods (WSS [34], ASMO [15], MWS [45], WSSA [48], MFNet [29] and SBBs [21]). For fair comparison, all the saliency maps are provided by the authors. In addition, our results are diametrically produce by CCFNet without relying on any post-processing.

Quantitative comparison The detailed F-measure, MAE, E-measure and S-measure values are provided in Table 2 and Table 3 on five common datasets, in which our approach performs favorably against other state-of-the-art unsupervised and weakly supervised approaches by a large margin, and even superior to some fully supervised methods, like Amulet, UCF and NLDF. It is worth noting that we also achieve the best results for saliency detection using challenging datasets, such as DUT-OMRON and DUTS-TE. For fairness, in terms of the average of each metric, we can conclude that our proposed approach shows a preferred average F_β (0.809 vs. 0.779), MAE (0.061 vs. 0.069), E_ϕ (0.880 vs. 0.875) and S_α (0.847 vs. 0.818) across five datasets than SBBs, which is the latest competitive algorithm. We have to admit that there are some gaps compared with the fully supervised algorithm in terms of performance even though it is reasonable and logical. Generally speaking, our approach is superior to other counterparts across all datasets using these evaluation metrics. Besides, Figure 8 and Figure 9 show the PR curves and F-measure curves with other weakly supervised state-of-the-art methods on the five benchmark datasets, respectively. It can be observed that our method achieves a better performance than the other ones in most cases. To further analyze the overall difference between our algorithm and other methods, we compare the quantitative results including average F-measure, average E-measure, average S-measure and average MAE on five common datasets, which are calculated by the average

Table 2 Comparison with other state-of-the-art approaches on ECSSD and DUT-OMRON datasets. 'F' means fully supervised, 'W' means weakly supervised and 'Un' is for unsupervised. ↑ & ↓ denote larger and smaller is better, respectively.

Methods	Year	Sup.	ECSSD 1000 images				DUT-OMRON 5168 images			
			$F_\beta \uparrow$	MAE ↓	$E_\phi \uparrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	MAE ↓	$E_\phi \uparrow$	$S_\alpha \uparrow$
Amulet [51]	ICCV 2017	F	0.868	0.059	0.901	0.894	0.647	0.098	0.779	0.781
UCF [52]	ICCV 2017	F	0.844	0.069	0.892	0.883	0.621	0.120	0.765	0.760
NLDF [24]	CVPR 2017	F	0.878	0.063	0.910	0.875	0.684	0.080	0.816	0.770
RAS [3]	ECCV 2018	F	0.889	0.056	0.914	0.893	0.713	0.062	0.846	0.814
PAGR[54]	CVPR 2018	F	0.894	0.061	0.914	0.889	0.711	0.071	0.842	0.775
BMPM [50]	CVPR 2018	F	0.868	0.045	0.914	0.911	0.692	0.064	0.837	0.809
DSS [12]	TPAMI 2019	F	0.904	0.052	0.912	0.882	0.740	0.063	0.842	0.790
EGNet [55]	ICCV 2019	F	0.920	0.037	0.927	0.925	0.755	0.053	0.868	0.841
CPD [40]	CVPR 2019	F	0.917	0.037	0.925	0.918	0.747	0.056	0.866	0.825
MINet [28]	CVPR 2020	F	0.924	0.033	0.927	0.925	0.755	0.056	0.865	0.833
VST [20]	ICCV 2021	F	0.920	0.033	0.918	0.932	0.756	0.058	0.861	0.850
SVF [47]	ICCV 2017	Un	0.809	0.088	0.875	0.832	0.608	0.108	0.768	0.747
C2S [18]	ECCV 2018	Un	0.853	0.059	0.906	0.882	0.664	0.079	0.817	0.780
WSS [34]	CVPR 2017	W	0.823	0.104	0.869	0.811	0.603	0.109	0.768	0.725
ASMO [15]	AAAI 2018	W	0.798	0.110	0.853	0.802	0.622	0.101	0.776	0.752
MWS [45]	CVPR 2019	W	0.840	0.096	0.884	0.827	0.609	0.109	0.763	0.756
WSSA [48]	CVPR 2020	W	0.870	0.059	0.901	0.865	0.703	0.068	0.840	0.785
MFNet [29]	ICCV 2021	W	0.844	0.084	0.877	0.837	0.621	0.098	0.783	0.726
SBBs [21]	TIP 2021	W	0.855	0.072	0.894	0.851	0.695	0.074	0.835	0.776
Ours	–	W	**0.890**	**0.050**	**0.912**	**0.882**	**0.720**	**0.069**	**0.848**	**0.796**

Table 3 Comparison with other state-of-the-art approaches on PASCAL-S, HKU-IS and DUTS-TE datasets. 'F' means fully supervised, 'W' means weakly supervised and 'Un' is for unsupervised. ↑ & ↓ denote larger and smaller is better, respectively. "-" means the authors did not release the code, and they just provided the saliency maps, thus reporting the total number of parameters of this method is not possible.

Methods	Year	Sup.	PASCAL-S 850 images				HKU-IS 4447 images				DUTS-TE 5019 images			
			$F_\beta \uparrow$	MAE ↓	$E_\phi \uparrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	MAE ↓	$E_\phi \uparrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	MAE ↓	$E_\phi \uparrow$	$S_\alpha \uparrow$
Amulet [51]	ICCV 2017	F	0.757	0.100	0.802	0.818	0.841	0.051	0.912	0.886	0.678	0.085	0.794	0.804
UCF [52]	ICCV 2017	F	0.726	0.115	0.804	0.805	0.823	0.062	0.902	0.875	0.631	0.112	0.763	0.782
NLDF [24]	CVPR 2017	F	0.769	0.098	0.839	0.805	0.874	0.048	0.929	0.887	-	-	-	-
RAS [3]	ECCV 2018	F	0.777	0.101	0.836	0.799	0.871	0.045	0.929	0.887	0.751	0.059	0.861	0.839
PAGR[54]	CVPR 2018	F	0.798	0.089	0.853	0.822	0.886	0.048	0.939	0.887	0.784	0.056	0.880	0.838
BMPM [50]	CVPR 2018	F	0.758	0.074	0.842	0.845	0.871	0.039	0.937	0.907	0.745	0.049	0.860	0.862
DSS [12]	TPAMI 2019	F	0.801	0.093	0.847	0.798	0.902	0.040	0.934	0.878	-	-	-	-
EGNet [55]	ICCV 2019	F	0.817	0.074	0.854	0.852	0.902	0.031	0.949	0.918	0.815	0.039	0.891	0.887
CPD [40]	CVPR 2019	F	0.820	0.071	0.855	0.848	0.891	0.034	0.944	0.905	0.805	0.043	0.886	0.869
MINet [28]	CVPR 2020	F	0.829	0.064	0.857	0.856	0.909	0.029	0.953	0.919	0.828	0.037	0.898	0.884
VST [20]	ICCV 2021	F	0.829	0.061	0.844	0.872	0.900	0.029	0.953	0.928	0.818	0.037	0.892	0.896
SVF [47]	ICCV 2017	Un	0.695	0.131	0.789	0.758	-	-	-	-	-	-	-	-
C2S [18]	ECCV 2018	Un	0.754	0.087	0.838	0.826	0.839	0.051	0.919	0.873	0.710	0.066	0.841	0.817
WSS [34]	CVPR 2017	W	0.715	0.139	0.791	0.744	0.821	0.079	0.896	0.822	0.654	0.100	0.795	0.748
ASMO [15]	AAAI 2018	W	0.693	0.149	0.772	0.717	0.806	0.086	0.878	0.804	0.614	0.116	0.772	0.697
MWS [45]	CVPR 2019	W	0.713	0.133	0.790	0.768	0.814	0.084	0.895	0.818	0.684	0.091	0.814	0.759
WSSA [48]	CVPR 2020	W	0.774	0.092	0.837	0.797	0.860	0.047	0.927	0.865	0.742	0.062	0.857	0.804
MFNet [29]	ICCV 2021	W	0.746	0.112	0.818	0.782	0.839	0.058	0.917	0.852	0.692	0.079	0.830	0.778
SBBs [21]	TIP 2021	W	-	-	-	-	0.843	0.056	0.920	0.854	0.722	0.073	0.851	0.789
Ours	–	W	**0.794**	**0.084**	**0.840**	**0.808**	**0.870**	**0.044**	**0.934**	**0.871**	**0.770**	**0.057**	**0.873**	**0.816**

scores for each metric. From Fig. 10, we observe that our method achieves the best performance in all four average metrics.

Qualitative comparison We also show some examples of saliency maps from our proposed model and other state-of-the-art methods using some challenging cases in Figure 11. For example, we use instances of large objects (1st and 2nd rows), multiple targets (3rd and 6th rows), complex scenes (4th and 7th rows), small objects (5th row), cluttered backgrounds (8th row) and low contrast (9th row). Specifically, the 1st shows a woman in an image and almost all methods are un-

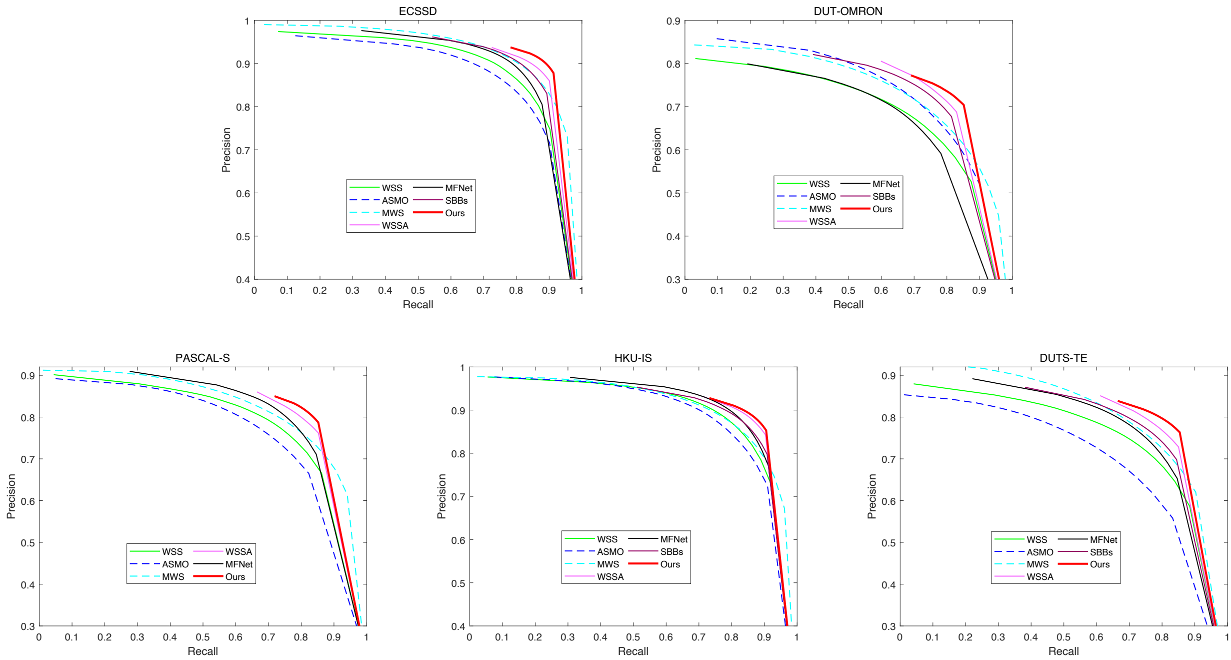

Fig. 8 PR curves of the proposed approach with other state-of-the-art methods using five datasets. Best viewed on screen.

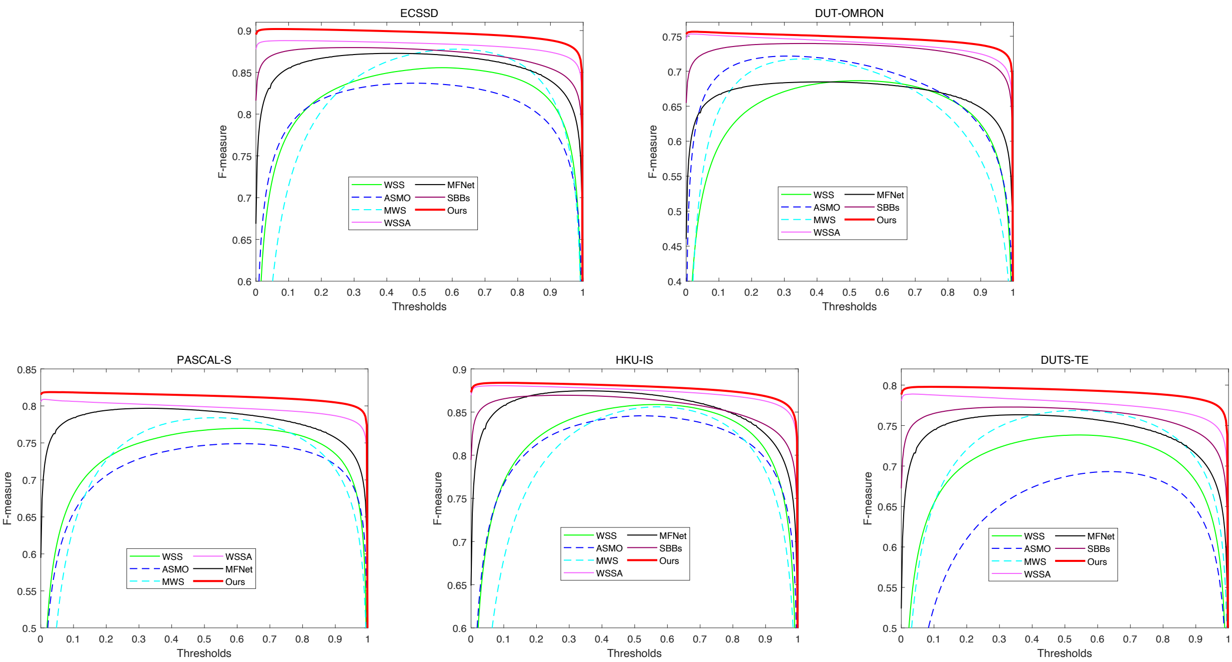

Fig. 9 F-measure curves of the proposed approach with other state-of-the-art methods using five datasets. Best viewed on screen.

able to detect accurate location on large objects except our method. In the 3rd row, most methods can locate the flowers while some details are lost. As we can see, our approach is able to accurately find the salient object with fewer false salient pixels detected. The 4th row corresponds to a dog in a stadium. It is easy to see our model segments the objects well, while other models always detect the alphabet as salient objects.

Fig. 10 (a)Comparison of quantitative results including average F-measure, average E-measure and average S-measure. Best viewed on screen. (b)Comparison of quantitative results including average MAE.

Compared with the 1st and 2nd row, a tower presented in the 5th row is more difficult to segment thanks to small objects and complex scenes. Nevertheless, our CCFNet still highlights it very well. Different from the last multiple examples, there are various characteristics of salient objects in the 6th row of Fig. 11, such as diverse colors and sizes. The proposed method can generate more reliable saliency maps in spite of the existing little deficiency. Although our CCFNet erroneously segments the bottom part, it is still much better than other methods. The 8th row shows the result of an object in cluttered backgrounds. Benefiting from EFM, our model has more accurate edge details. Furthermore, the 9th row demonstrates that our model has good performance with low contrast between the target and image background. It can be observed that our model is able to produce the complete structure of the cup whereas previous work can not. In conclusion, our proposed approach performs better with respect to salient object segmentation and localization, generating results that are much closer to the ground truth in various challenging scenarios.

4.4 Ablation studies

In this section, we perform a series of cases on ECSSD and DUTS-TE datasets to assess the effectiveness of our proposed method. All the ablation studies follow the same implementation setup.

Validity of different proposed module We conduct various experiments to verify the effectiveness of each component in CCFNet. In order to prove the validity of the proposed modules for saliency detection, we compare our method with the other six schemes with different proposed modules. Table 4 shows the performance with seven schemes as well as their corresponding saliency detection results. As seen from this table, on the one hand, the quantitative scores of 1st $\sim$ 3rd lines are lower than 4th $\sim$ 6th lines, that is both two modules added is superior to single module, meanwhile, the quantitative scores of 4th $\sim$ 6th lines are lower than 7th line (ours), that is to say, three modules work together to realize the significant results. Especially, we observe that these results perform more obviously on complex datasets, such as DUTS-TE. Furthermore, Fig. 12 shows the results of average F-measure, average E-measure, average S-measure and average MAE on ECSSD and DUTS-TE datasets. It also can be seen that the last scheme achieves the best performance, i.e. the scheme adopts three modules simultaneously.

Validity of different loss functions There are three types of key loss function within the CCFNet, i.e., PCE loss, LSC loss and SSD loss. We design three ablation experiments to evaluate the necessity of each loss function, F-measure, MAE, E-measure and S-measure scores are shown in Table 5. We find that LSC loss and SSD loss can boost the performance of saliency maps based on only using PCE loss. Especially, compared with the first line (w/o LSC & SSD loss), our CCFNet would promote the final performance with about 17.7%, 45.1%, 9.7% and 2.9% in F-measure, MAE, E-measure and S-measure scores on ECSSD datasets, respectively. In addition, Fig. 13 illustrates some samples of the different loss functions, which can be seen that the proposed CCFNet is well applicable to single target (1st and 2nd lines) or multiple targets (3rd line).

Validity of different parameters of loss function Here we analyze the Validity of different parameters of the loss function in Table 6. Accordingly, $\gamma_1 = 0, \gamma_2 = 0, \gamma_3 = 0$ means that our network has no extra supervision except dominant loss function L_{dom}. It can be seen that it has the lowest scores compared with

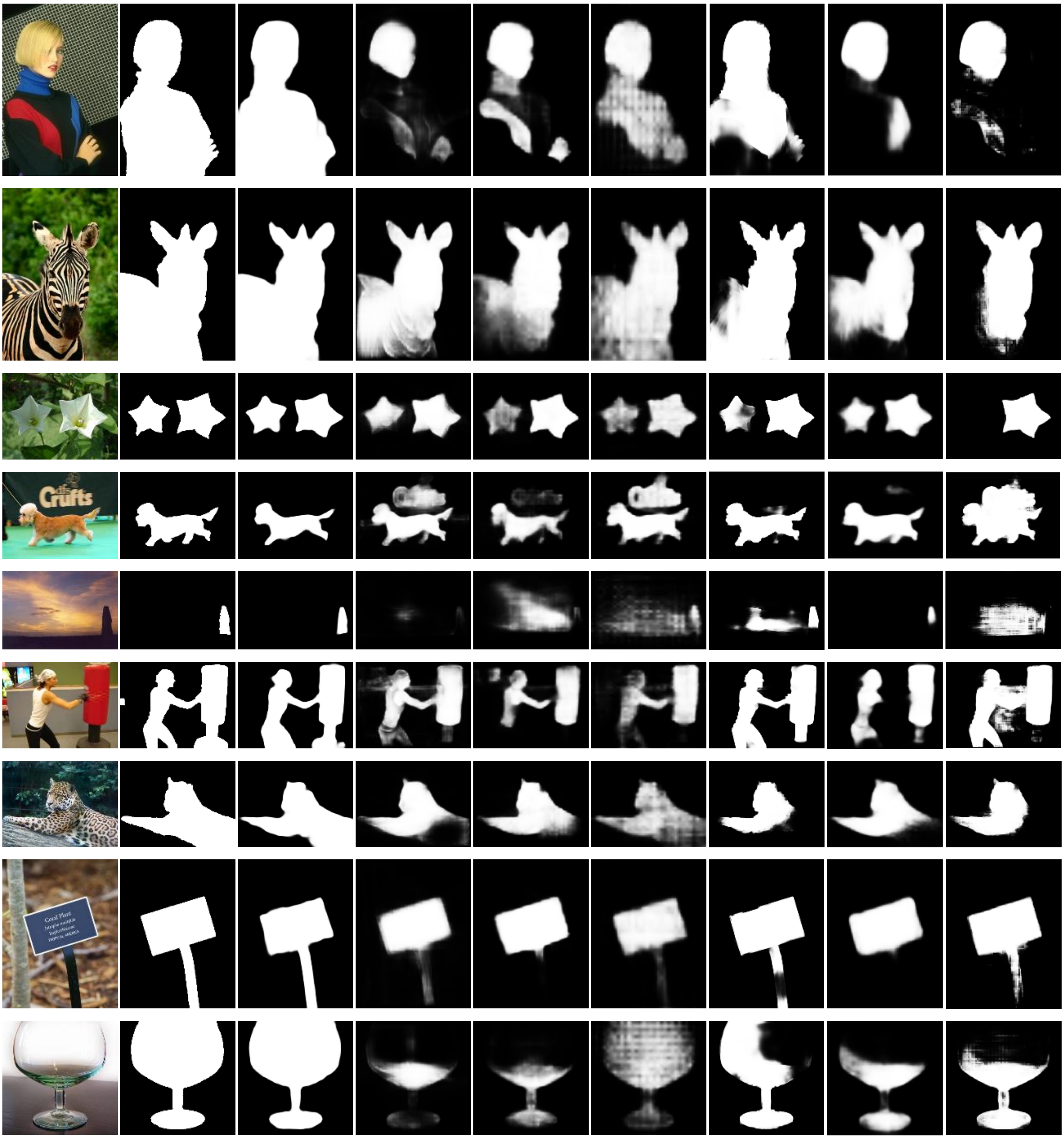

Fig. 11 Visual comparison between the proposed model and state-of-the-art methods.

other schemes. That means sub-stage loss is beneficial to the network. Similarly, $\gamma_1 = 1, \gamma_2 = 1, \gamma_3 = 1$ means that sub-stage loss has the same weight with the dominant loss function L_{dom}. This is not the best result for the scribble saliency detection network, which may be caused by the sub-stage bringing more negative information. Just because there is more and more useful information from the first sub-stage to the third sub-stage, hence, we balance the weight coefficients in a progressive manner. It is proved that our approach is effective and reliable. Note that $\gamma_1 = 0.8, \gamma_2 = 0, \gamma_3 = 0.4$ means that there is no first sub-stage loss function,

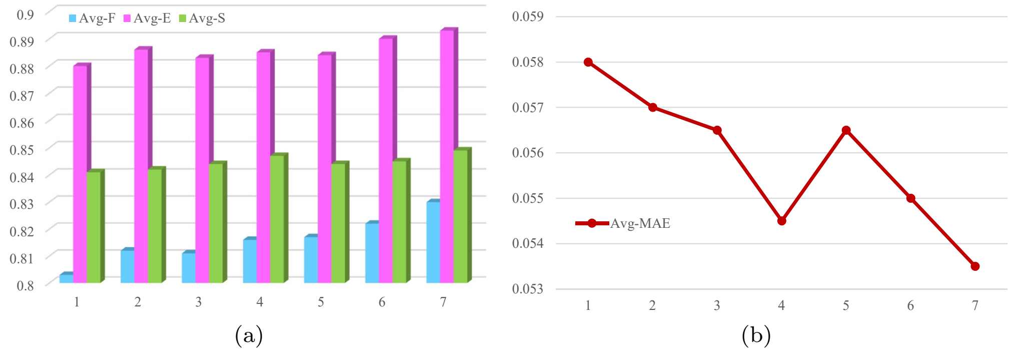

Fig. 12 (a)Comparison of quantitative results including average F-measure, average E-measure and average S-measure. Best viewed on screen. (b)Comparison of quantitative results including average MAE.

Table 4 Ablation study for our proposed different modules.

				ECSSD				DUTS-TE			
	FCM	EFM	GCGO	$F_\beta \uparrow$	MAE $\downarrow$	$E_\phi \uparrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	MAE $\downarrow$	$E_\phi \uparrow$	$S_\alpha \uparrow$
1			✓	0.869	0.054	0.905	0.874	0.737	0.062	0.854	0.807
2	✓			0.874	0.054	0.907	0.874	0.749	0.060	0.864	0.810
3		✓		0.874	0.052	0.907	0.877	0.747	0.061	0.859	0.811
4		✓	✓	0.878	0.051	0.905	0.879	0.753	0.058	0.864	0.815
5	✓		✓	0.880	0.053	0.904	0.876	0.754	0.060	0.864	0.812
6	✓	✓		0.881	0.052	0.910	0.877	0.762	0.058	0.870	0.813
7	✓	✓	✓	**0.890**	**0.050**	**0.912**	**0.882**	**0.770**	**0.057**	**0.873**	**0.816**

Table 5 Ablation study for different loss functions.

				ECSSD				DUTS-TE			
	PCE	LSC	SSD	$F_\beta \uparrow$	MAE $\downarrow$	$E_\phi \uparrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	MAE $\downarrow$	$E_\phi \uparrow$	$S_\alpha \uparrow$
1	✓			0.756	0.091	0.831	0.799	0.575	0.105	0.725	0.711
2	✓		✓	0.763	0.087	0.839	0.809	0.589	0.095	0.742	0.727
3	✓	✓		0.875	0.053	0.900	0.875	0.751	0.052	0.857	0.812
4	✓	✓	✓	**0.890**	**0.050**	**0.912**	**0.882**	**0.770**	**0.057**	**0.873**	**0.816**

Table 6 Ablation study for different parameters of loss function.

		ECSSD				DUTS-TE			
		$F_\beta \uparrow$	MAE $\downarrow$	$E_\phi \uparrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	MAE $\downarrow$	$E_\phi \uparrow$	$S_\alpha \uparrow$
1	$\gamma_1 = 0, \gamma_2 = 0, \gamma_3 = 0$	0.866	0.055	0.896	0.872	0.732	0.064	0.852	0.804
2	$\gamma_1 = 0, \gamma_2 = 0.6, \gamma_3 = 0.4$	0.869	0.054	0.901	0.874	0.732	0.064	0.848	0.807
3	$\gamma_1 = 0.8, \gamma_2 = 0, \gamma_3 = 0.4$	0.878	0.051	0.911	0.878	0.758	0.057	0.873	0.815
4	$\gamma_1 = 0.8, \gamma_2 = 0.6, \gamma_3 = 0$	0.874	0.051	0.906	0.879	0.734	0.065	0.848	0.807
5	$\gamma_1 = 1, \gamma_2 = 1, \gamma_3 = 1$	0.875	0.052	0.900	0.876	0.747	0.061	0.856	0.811
6	$\gamma_1 = 0.8, \gamma_2 = 0.6, \gamma_3 = 0.4$	**0.890**	**0.050**	**0.912**	**0.882**	**0.770**	**0.057**	**0.873**	**0.816**

which can be seen that the result is poor on two datasets. The reason for this issue may be that the first substage supervision has more semantics, which plays a decisive role in the subsequent prediction.

5 Conclusion

In this paper, we proposed a novel and effective complementary characteristics fusion network (CCFNet) for salient object detection with scribble annotations. First, a global context guiding operation and edge fusion mod-

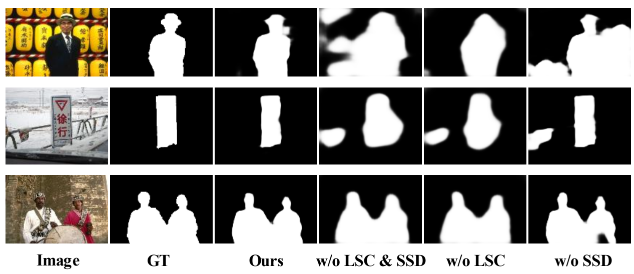

Fig. 13 Ablation study of different loss function.

ule are introduced, which are used to obtain global semantics and learn salient edge information. It is proved that they can better understand global high level information and edge information. Next, to exploit complete salient regions with different level features, this paper proposes the feature correlation module for saliency detection. In order to better distinguish foreground and background information for a given image, self-supervised saliency detection loss is illustrated. Finally, to demonstrate the performance of our proposed method, we conduct experiment results on five well-known datasets. Extensive experimental results demonstrate that our approach outperforms state-of-the-art weakly supervised methods and ablation studies prove the effectiveness of each component as well. Future work will focus on developing a more lightweight weakly supervised model as well as investigating how to deploy SOD algorithms in mobile devices to strengthen practicality.

Declaration of Interests

We declare that we have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by National Natural Science Foundation of China (No. 61973066,61471110) , Major Science and Technology Projects of Liaoning Province (No.2021JH1/10400049), Fundation of Key Laboratory of Aerospace System Simulation(No.6142002200301), Open Research Projects of Zhejiang Lab (No.2019KD0AD01/006) and Major Science and technology innovation engineering projects of Shandong Province (No.2019JZZY010128).

References

1. Araslanov, N., Roth, S.: Single-stage semantic segmentation from image labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4253–4262 (2020)
2. Borji, A., Itti, L.: Exploiting local and global patch rarities for saliency detection. In: 2012 IEEE conference on computer vision and pattern recognition, pp. 478–485. IEEE (2012)
3. Chen, S., Tan, X., Wang, B., Hu, X.: Reverse attention for salient object detection. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 234–250 (2018)
4. Cheng, M.M., Mitra, N.J., Huang, X., Torr, P.H., Hu, S.M.: Global contrast based salient region detection. IEEE transactions on pattern analysis and machine intelligence **37**(3), 569–582 (2014)
5. Cheng, M.M., Zhang, F.L., Mitra, N.J., Huang, X., Hu, S.M.: Repfinder: finding approximately repeated scene elements for image editing. ACM Transactions on Graphics (TOG) **29**(4), 1–8 (2010)
6. Craye, C., Filliat, D., Goudou, J.F.: Environment exploration for object-based visual saliency learning. In: 2016 IEEE International Conference on Robotics and Automation (ICRA), pp. 2303–2309. IEEE (2016)
7. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International journal of computer vision **88**(2), 303–338 (2010)
8. Fan, D.P., Cheng, M.M., Liu, Y., Li, T., Borji, A.: Structure-measure: A new way to evaluate foreground maps. In: Proceedings of the IEEE international conference on computer vision, pp. 4548–4557 (2017)
9. Fan, D.P., Gong, C., Cao, Y., Ren, B., Cheng, M.M., Borji, A.: Enhanced-alignment measure for binary foreground map evaluation. arXiv preprint arXiv:1805.10421 (2018)
10. Goferman, S., Zelnik-Manor, L., Tal, A.: Context-aware saliency detection. IEEE transactions on pattern analysis and machine intelligence **34**(10), 1915–1926 (2011)
11. Hong, S., You, T., Kwak, S., Han, B.: Online tracking by learning discriminative saliency map with convolutional neural network. In: International conference on machine learning, pp. 597–606. PMLR (2015)
12. Hou, Q., Cheng, M.M., Hu, X., Borji, A., Tu, Z., Torr, P.: Deeply supervised salient object detection with short connections. IEEE Transactions on Pattern Analysis and Machine Intelligence **41**(4), 815–828 (2019)
13. Hou, Q., Cheng, M.M., Hu, X., Borji, A., Tu, Z., Torr, P.H.: Deeply supervised salient object detection with short connections. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3203–3212 (2017)
14. Jiang, H., Wang, J., Yuan, Z., Wu, Y., Zheng, N., Li, S.: Salient object detection: A discriminative regional feature integration approach. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2083–2090 (2013)
15. Li, G., Xie, Y., Lin, L.: Weakly supervised salient object detection using image labels. In: Thirty-second AAAI conference on artificial intelligence (2018)
16. Li, G., Yu, Y.: Visual saliency based on multiscale deep features. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5455–5463 (2015)
17. Li, G., Yu, Y.: Visual saliency detection based on multiscale deep cnn features. IEEE transactions on image processing **25**(11), 5012–5024 (2016)
18. Li, X., Yang, F., Cheng, H., Liu, W., Shen, D.: Contour knowledge transfer for salient object detection. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 355–370 (2018)
19. Liu, J.J., Hou, Q., Cheng, M.M., Feng, J., Jiang, J.: A simple pooling-based design for real-time salient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3917–3926 (2019)
20. Liu, N., Zhang, N., Wan, K., Han, J., Shao, L.: Visual saliency transformer. arXiv preprint arXiv:2104.12099 (2021)
21. Liu, Y., Wang, P., Cao, Y., Liang, Z., Lau, R.W.: Weakly-supervised salient object detection with saliency bounding boxes. IEEE Transactions on Image Processing **30**, 4423–4435 (2021)

22. Liu, Y., Zhang, Y., Liu, S., Coleman, S., Wang, Z., Qiu, F.: Salient object detection by aggregating contextual information. Pattern Recognition Letters **153**, 190–199 (2022)
23. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3431–3440 (2015)
24. Luo, Z., Mishra, A., Achkar, A., Eichel, J., Li, S., Jodoin, P.M.: Non-local deep features for salient object detection. In: Proceedings of the IEEE Conference on computer vision and pattern recognition, pp. 6609–6617 (2017)
25. Nguyen, D.T., Dax, M., Mummadi, C.K., Ngo, T.P.N., Nguyen, T.H.P., Lou, Z., Brox, T.: Deepusps: Deep robust unsupervised saliency prediction with self-supervision. arXiv preprint arXiv:1909.13055 (2019)
26. Noori, M., Mohammadi, S., Majelan, S.G., Bahri, A., Havaei, M.: Dfnet: Discriminative feature extraction and integration network for salient object detection. Engineering Applications of Artificial Intelligence **89**, 103419 (2020)
27. Pan, Z., Jiang, P., Tu, C.: Scribble-supervised semantic segmentation by random walk on neural representation and self-supervision on neural eigenspace. arXiv preprint arXiv:2011.05621 (2020)
28. Pang, Y., Zhao, X., Zhang, L., Lu, H.: Multi-scale interactive network for salient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9413–9422 (2020)
29. Piao, Y., Wang, J., Zhang, M., Lu, H.: Mfnet: Multi-filter directive network for weakly supervised salient object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4136–4145 (2021)
30. Qin, X., Zhang, Z., Huang, C., Dehghan, M., Zaiane, O.R., Jagersand, M.: U2-net: Going deeper with nested u-structure for salient object detection. Pattern Recognition **106**, 107404 (2020)
31. Siva, P., Russell, C., Xiang, T., Agapito, L.: Looking beyond the image: Unsupervised learning for object saliency and detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3238–3245 (2013)
32. Tang, M., Djelouah, A., Perazzi, F., Boykov, Y., Schroers, C.: Normalized cut loss for weakly-supervised cnn segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1818–1827 (2018)
33. Ullah, I., Jian, M., Hussain, S., Lian, L., Ali, Z., Qureshi, I., Guo, J., Yin, Y.: Global context-aware multi-scale features aggregative network for salient object detection. Neurocomputing **455**, 139–153 (2021)
34. Wang, L., Lu, H., Wang, Y., Feng, M., Wang, D., Yin, B., Ruan, X.: Learning to detect salient objects with image-level supervision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 136–145 (2017)
35. Wang, T., Borji, A., Zhang, L., Zhang, P., Lu, H.: A stagewise refinement model for detecting salient objects in images. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4019–4028 (2017)
36. Wang, W., Shen, J., Dong, X., Borji, A.: Salient object detection driven by fixation prediction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1711–1720 (2018)
37. Wei, J., Wang, S., Huang, Q.: F^3net: Fusion, feedback and focus for salient object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 12321–12328 (2020)
38. Wei, J., Wang, S., Wu, Z., Su, C., Huang, Q., Tian, Q.: Label decoupling framework for salient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13025–13034 (2020)
39. Wei, Y., Liang, X., Chen, Y., Shen, X., Cheng, M.M., Feng, J., Zhao, Y., Yan, S.: Stc: A simple to complex framework for weakly-supervised semantic segmentation. IEEE transactions on pattern analysis and machine intelligence **39**(11), 2314–2320 (2016)
40. Wu, Z., Su, L., Huang, Q.: Cascaded partial decoder for fast and accurate salient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3907–3916 (2019)
41. Wu, Z., Su, L., Huang, Q.: Stacked cross refinement network for edge-aware salient object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7264–7273 (2019)
42. Yan, Q., Xu, L., Shi, J., Jia, J.: Hierarchical saliency detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1155–1162 (2013)
43. Yang, C., Zhang, L., Lu, H., Ruan, X., Yang, M.H.: Saliency detection via graph-based manifold ranking. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3166–3173 (2013)
44. Yu, S., Zhang, B., Xiao, J., Lim, E.G.: Structure-consistent weakly supervised salient object detection with local saliency coherence. In: AAAI Conf. Art. Intell (2021)
45. Zeng, Y., Zhuge, Y., Lu, H., Zhang, L., Qian, M., Yu, Y.: Multi-source weak supervision for saliency detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6074–6083 (2019)
46. Zhai, Y., Shah, M.: Visual attention detection in video sequences using spatiotemporal cues. In: Proceedings of the 14th ACM international conference on Multimedia, pp. 815–824 (2006)
47. Zhang, D., Han, J., Zhang, Y.: Supervision by fusion: Towards unsupervised learning of deep salient object detector. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4048–4056 (2017)
48. Zhang, J., Yu, X., Li, A., Song, P., Liu, B., Dai, Y.: Weakly-supervised salient object detection via scribble annotations. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 12546–12555 (2020)
49. Zhang, J., Zhang, T., Dai, Y., Harandi, M., Hartley, R.: Deep unsupervised saliency detection: A multiple noisy labeling perspective. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 9029–9038 (2018)
50. Zhang, L., Dai, J., Lu, H., He, Y., Wang, G.: A bi-directional message passing model for salient object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1741–1750 (2018)
51. Zhang, P., Wang, D., Lu, H., Wang, H., Ruan, X.: Amulet: Aggregating multi-level convolutional features for salient object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 202–211 (2017)
52. Zhang, P., Wang, D., Lu, H., Wang, H., Yin, B.: Learning uncertain convolutional features for accurate saliency detection. In: Proceedings of the IEEE International Conference on computer vision, pp. 212–221 (2017)

53. Zhang, P., Zhuo, T., Huang, W., Chen, K., Kankanhalli, M.: Online object tracking based on cnn with spatial-temporal saliency guided sampling. Neurocomputing **257**, 115–127 (2017)
54. Zhang, X., Wang, T., Qi, J., Lu, H., Wang, G.: Progressive attention guided recurrent network for salient object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 714–722 (2018)
55. Zhao, J.X., Liu, J.J., Fan, D.P., Cao, Y., Yang, J., Cheng, M.M.: Egnet: Edge guidance network for salient object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 8779–8788 (2019)
56. Zhao, T., Wu, X.: Pyramid feature attention network for saliency detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019)

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: