



A novel seminar learning framework for weakly supervised salient object detection

Liu, Y., Zhang, Y., Wang, Z., Yang, F., Qiu, F., Coleman, S., & Kerr, D. (2023). A novel seminar learning framework for weakly supervised salient object detection. *Engineering Applications of Artificial Intelligence*, 126(Part B), 1-17. Article 106961. <https://doi.org/10.1016/j.engappai.2023.106961>

[Link to publication record in Ulster University Research Portal](#)

Published in:
Engineering Applications of Artificial Intelligence

Publication Status:
Published (in print/issue): 30/11/2023

DOI:
[10.1016/j.engappai.2023.106961](https://doi.org/10.1016/j.engappai.2023.106961)

Document Version
Author Accepted version

Document Licence:
CC BY-NC-ND

General rights

The copyright and moral rights to the output are retained by the output author(s), unless otherwise stated by the document licence.

Unless otherwise stated, users are permitted to download a copy of the output for personal study or non-commercial research and are permitted to freely distribute the URL of the output. They are not permitted to alter, reproduce, distribute or make any commercial use of the output without obtaining the permission of the author(s).

If the document is licenced under Creative Commons, the rights of users of the documents can be found at <https://creativecommons.org/share-your-work/cclicenses/>.

Take down policy

The Research Portal is Ulster University's institutional repository that provides access to Ulster's research outputs. Every effort has been made to ensure that content in the Research Portal does not infringe any person's rights, or applicable UK laws. If you discover content in the Research Portal that you believe breaches copyright or violates any law, please contact pure-support@ulster.ac.uk

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

A Novel Seminar Learning Framework for Weakly Supervised Salient Object Detection

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

Received: date / Accepted: date

Abstract Weakly supervised salient object detection (SOD) is a challenging task and has drawn much attention from several research perspectives, it has revealed two problems while driving the rapid development of saliency detection. (1) Large divergence in the characteristics of saliency regions in terms of location, shape and size makes them difficult to recognize. (2) The properties of convolutional neural networks dictate that it is insensitive to various transformations, which will lead to hardly balance the application of various disturbances. To tackle these limitations, this paper proposes a novel seminar learning framework with consistent transformation ensembling (SLF-CT) for scribble supervised SOD. The framework consists of the teacher-student model and the student-student model for segmenting the salient objects. Specifically, we first design a cross attention guided network (CAGNet) as a baseline model for saliency prediction. Then we assign CAGNet to the teacher-student model, where the teacher network is based on the exponential moving average and guides the training of the student network. Moreover, we adopt multiple pseudo labels to transfer the information among students from different conditions. To further enhance the regularization of the network, a consistency transformation mechanism is also incorporated, which encourages the saliency prediction and

input image of the network to be consistent. The experimental results demonstrate that the proposed approach performs favorably comparable with the state-of-the-art weakly supervised methods. As far as we know, the proposed approach is the first application of seminar learning in the SOD area.

Keywords Salient object detection, [weakly supervised](#), [seminar learning framework](#), [cross attention guided network](#), [consistency transformation mechanism](#), [multiple pseudo labels](#)

1 Introduction

Salient object detection (SOD) aims to locate and segment the most dominant objects from an input image [26, 46, 81, 38, 27, 9, 16]. It plays a significant role in a variety of computer vision related fields, such as scene classification [51], object segmentation [61], video object tracking [19] and robot navigation [11]. Although the existing approaches [35, 33, 31, 46] achieve remarkable performance, they strongly rely on a large amount of accurately labeled data with full pixel-wise annotations for training, which takes extensive effort to collect. Once the annotation problem is addressed, more accurate prediction among images can be achieved. Therefore, it is worth investigating how to leverage sparse labels to boost the accuracy of saliency detection. To balance the trade-off between label efficiency and model performance, some methods based on sparse labels have been proposed [55, 39, 74], and in particularly, scribble annotation has proven to be a valuable approach and is the focus of this work.

Saliency detection with scribble annotation [focuses on learning](#) from minimal annotation, this is a fundamental and challenging task. Recent progress in weakly

* Corresponding author.

Y. Liu, and Z. Wang are with Faculty of Robot Science and Engineering, Northeastern University, Shenyang, China.

Y. Zhang 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.

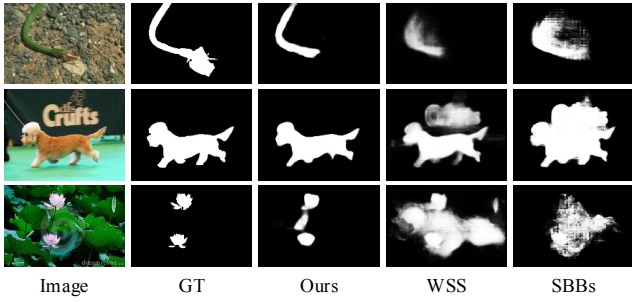


Fig. 1 Comparison between WSS [55], SBBs [39] and our SLF-CT algorithm.

supervised learning for saliency detection has featured deep learning approaches. A scribble boosting scheme was proposed in work [74] to train the model with initial saliency predictions, which are used for obtaining improved saliency maps. Wang et al. [55] showed an image-level supervised learning model for saliency detection, where the model leverages the foreground inference network to learn salient object detectors. Different from earlier work [74, 55], we tackle the weakly supervised SOD task from a new perspective, i.e. using a seminar learning strategy.

Although advances have been made via previous methods, there still remain two key problems. First, large divergence in the characteristics of saliency regions in terms of location, shape and size makes them difficult to recognize, especially for weakly supervised learning methods, for example, see Fig. 1. The limited information that is provided by scribble annotations, which do not cover whole objects results in defective predictions. As we know, the network trained under different conditions lead to different prediction. This uncertainty has led to different information generated in different network, which can be gathered together to complement each other. The teacher-student model [6], has benefited some works [37, 12, 50] to solve the above conditions and they have also demonstrated improved performance. For instance, [37] designed an uncertainty-aware mean teacher architecture to guide the student model dealing with various infection regions in medical related images. The work in [12] also introduced a teacher-student model to alleviate the label scarcity problem in a semi-supervised learning way. However, these methods only consider the relationship between the teacher and the student, excluding other situations. Inspired by what happens in a real school, that is the student not only can learn from the teacher, but also can learn from the other students, we propose the seminar learning strategy in this work.

Second, owing to the CNNs being less sensitive to some spatial transformations (e.g. positional variation),

applying transformations to obtain a good saliency map is difficult. Some related works [36, 49] designed a self-ensembling (SE) framework with transformations to classify medical images or perform object detection. It is well known that if a given image is transformed (e.g. rotation), the result will be transformed in the same way. This has been demonstrated for classification problems. Nevertheless, this advantage is not preserved for SOD due to the transformation equivariant. With the success of the SE framework in the semi-supervised tasks, we propose to take this favorable characteristic into weakly supervised salient object detection tasks. In other words, to ease the limitation of transformation, we employ a consistency transformation strategy to balance the application of various disturbances to the input of the model, and can learn more different forms of the same salient target.

Based on the above considerations, this paper proposes a seminar learning framework with consistent transformation ensembling (SLF-CT) for weakly supervised SOD, including the ancillary stage and the primary stage. The ancillary stage is responsible for generating various pseudo labels, the primary stage is the final training stage under different conditions. Since seminar learning is a strategy but itself doesn't contain any network, thus we have to design the network for salient object detection. At the same time, considering that scribble annotations are sparse label, we need to extract and integrate effective information in the network as much as possible. Motivated by these, we first present a cross attention guided network (CAGNet) to direct the model predictions. Followed by this, we utilize CAGNet as both the teacher network and the student network in different stages (i.e. ancillary and primary stage). The teacher network is updated by the exponential moving average (EMA) algorithm of the student network to ensemble the information in different training steps. Alternatively, we present a seminar learning framework on how to use the student-student and teacher-student models to boost the performance of saliency maps detection. With the help of seminar learning way, a student (primary student) is able to learn from both the teacher (primary teacher) and another student (ancillary student). It is worth mentioning that we integrate a consistent transformation mechanism in the teacher-student model, which highly guarantees invariant predictions for perturbations of scribble annotation. Despite work [7] also adopting seminar learning style by click-level supervision, they neglect the role of consistency transformation. With respect to the student-student model, we apply multiple pseudo labels to enhance the efficiency of knowledge transfer.

From the above, our main contributions can be summarized as follows:

- We develop a seminar learning framework with consistent transformation ensembling (SLF-CT) for weakly supervised saliency prediction. To the best of our knowledge, this is the first weakly supervised SOD approach by using seminar learning.
- We regularize the model with a consistent transformation mechanism, which encourages the saliency prediction to be consistent under various spatial transformations.
- Multiple pseudo labels are applied to the student-student model that is used to strengthen the efficiency of different information transfer, which can complement each other.
- Experimental results show that our proposed SLF-CT outperforms the existing state-of-the-art methods on five public datasets, even surpasses the performance of those fully supervised methods.

2 Related Work

In this section, we briefly review some related works, including salient object detection and the consistent transformation mechanism. The salient object detection methods include fully supervised, weakly supervised and unsupervised methods.

Fully supervised salient object detection Traditional SOD approaches mainly depend on some hand-crafted features [3, 10, 1, 72] to directly detect salient objects, such as color difference [1] and contrast [10]. While these approaches are limited to representation capacity, deep learning based methods [5, 21, 4, 45, 65] have become popular due to their accurate performance. A variety of effective fully convolutional network based (FCN-based) architectures [40, 60, 56] have been proposed to enhance the generation of saliency maps in literature. For instance, in [56], 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 F³Net was introduced in [60], to solve the problem generated by the different receptive fields of different convolutional layers. Liu et al. [40] have proposed an aggregating contextual information network, which fully takes advantage of both low-level texture information and high-level semantic information. Nevertheless, these methods only concentrate on the issue domain within the scope of the single image. Based on this consideration, Ma et al. [42] attempted to adopt those beyond-scope information to promote the current SOD. Recently, attention mechanism has at-

tracted interest in saliency prediction [8, 79, 81]. Chen et al. [8] designed reverse attention to learn residual details for saliency refinement. A number of related works [32, 80, 59, 57] also leveraged the edge information to learn more detailed features. For example, [80, 59] used edge ground-truth as auxiliary supervision, proving that it is helpful for saliency maps, especially with object boundaries. Different from the above, Wu et al. [62] developed the "dynamic" scale routing for saliency detection, which motivated by the kernels of different size are preferable different salient objects, e.g., kernels of small size are adaptive for tiny salient objects, kernels of large size are adaptive for large salient objects.

Weakly supervised salient object detection

Due to the restricted time and cost of labeling, lots of sparse labels are introduced for the saliency detection task. Typical weak annotations include noisy labels [75], bounding boxes [52, 39], scribble annotations [70, 74] and image-level labels [24, 55, 71], which have received a lot of attention from researchers. Wang et al. [55] adopted a foreground inference network for object saliency prediction with image-level labels, which was the first application of this to SOD. Subsequently, a multi-task fully convolutional network is introduced in [24] to generate saliency maps using image-level supervision. Scribble annotation is commonly applied to segmentation filed [30], whereas it has attracted less attention in saliency detection. In [74], the S-DUTS dataset was firstly proposed for SOD owing to their flexibility to label a variety of objects, being less time-consuming and its low cost compared with per-pixel masks. After that, Yu et al. [70] also used S-DUTS dataset to develop a one-round end-to-end approach with local saliency coherence for scribble supervised saliency detection. Currently, in [39], Liu et al. fused the unsupervised SOD methods and bounding boxes in the different training steps. A related work [47] built a saliency network and multiple directive filters to enhance the performance of SOD, which is a multiple-pseudo-label framework. In addition, some semi-supervised methods are also proposed for reducing the cost of labeling. Wu et al. [64] utilized the synthesized pairs replacing the human-labeled DUTS-TR dataset to train the SOD model. Similarly, Yan et al. [66] solved the semi-supervised video SOD task using pseudo-labels as well, which includes a spatial refinement network and a spatiotemporal module. Different from these methods, we adopt a novel approach to reduce the gap between fully supervised learning and weakly supervised learning, which is the first weakly supervised saliency detection approach by using the seminar learning method, it can even outperform some fully supervised methods.

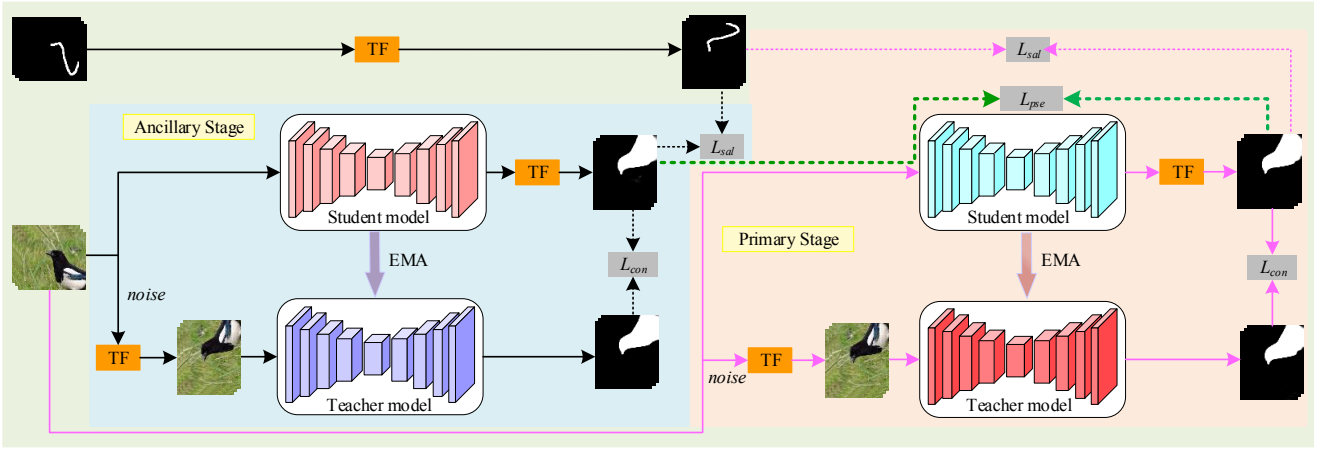


Fig. 2 Illustration of our proposed seminar learning framework with consistent transformation ensembling (SLF-CT) for scribble supervised salient object detection. SLF-CT contains of two stages in the training process, i.e. ancillary stage and primary stage. The teacher and student model are the same architecture in each stage, and the teacher model is updated by the EMA of the student model. "TF" means transformation operation.

Unsupervised salient object detection Early methods have been proposed for predicting the saliency map, which mainly used some handcrafted features, contrast, different priors etc., [17, 23, 10]. With the development of deep learning, Zhang et al. [73] proposed a deep learning framework from unsupervised methods with heuristics to produce saliency maps. Moreover, Li et al. [28] developed a contour-to-saliency network based on the existing well-trained contour detection network. Nguyen et al. [44] subsequently introduced an approach based on noisy label generation to improve prediction quality. Despite unsupervised SOD making great advances, the accuracy still needs to be improved.

Consistent transformation The consistent transformation has been utilized widely in sparse label based weakly supervised learning. It can improve the performance of classification and segmentation models when there is a large number of unannotated images and a limited number of annotations. [29, 37, 12, 36, 50]. Generally speaking, they are built by the teacher model and the student model. For example, Li et al. [29] proposed a transformation-consistent self-ensembling model (TCSM_v2) to improve the performance for the semi-supervised medical image segmentation. Furthermore, the uncertainty-aware self-ensembling and transformation-consistent mean teacher model (USTM) [37] was introduced with sparse annotations, which can encourage the segmentation predictions of medical images to be consistent under different perturbations for an input image. Liu et al. [36] introduced a sample-relation-consistent mean-teacher framework (SRC-MT). This method utilizes unlabeled data by encouraging predictive consistency for a given input under perturbations, and uses a self-ensembling model to produce high-quality

consistency targets for unlabeled data. Inspired by the success of the consistent transformation mechanisms in many computer vision tasks, we propose the first weakly supervised SOD approach by taking the consistent transformation mechanism as the basic seminar network. Differently, our method not only using the relationship between the teacher and the student, but also considering the student and the student.

3 Methodology

In this section, we will elaborate on the proposed seminar learning framework with consistent transformation ensembling (SLF-CT) for weakly supervised SOD. In Section 3.1, the overview is described. Then the seminar learning framework is presented in Section 3.2. Finally, we describe the loss function in Section 3.3.

3.1 Overview

The teacher-student model and student-student model used in our framework are similar to what happens in a real school and this is called seminar learning. Inspired by this, we propose a seminar learning framework with consistent transformation ensembling (SLF-CT) for weakly supervised SOD, as shown in Fig. 2. The training dataset is given as $D = \{(x_i, s_i)\}_{i=1}^M$, where $x_i \in X$ is the original image, $s_i \in S$ is the corresponding scribble annotation and M is the total number of training images. Our proposed SLF-CT comprises two stages, the ancillary stage and primary stage. The teacher and student models are the same architecture

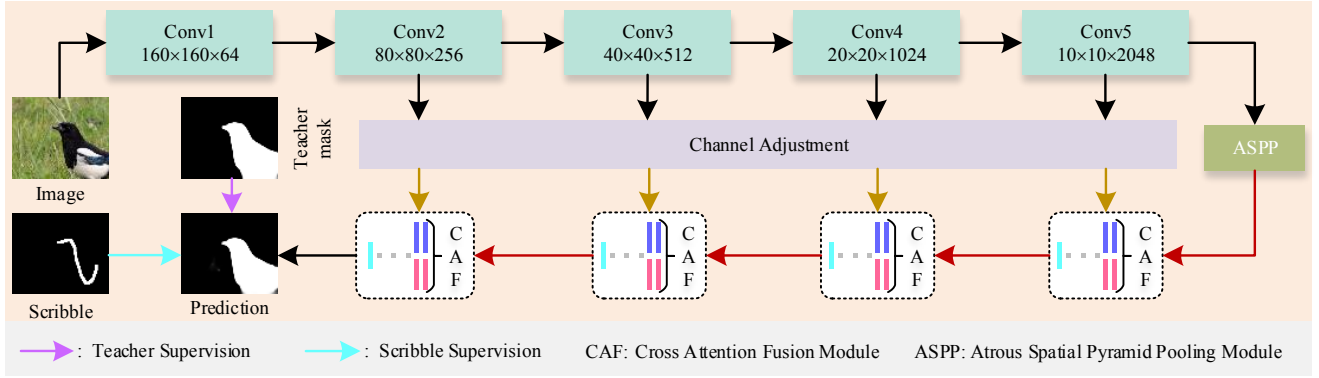


Fig. 3 Detailed structure of our cross attention guided network (CAGNet) with scribble annotations.

in each stage, and the parameters of the teacher network are updated by the EMA of the parameters of the corresponding student network.

For both ancillary and primary stages, we randomly extract images x_i from the original training image sets under noise conditions and put them into the different teacher network, respectively. Meanwhile, the original images are transformed (including random rotation, random flipping and random scaling) before entering the teacher network. For consistency, the same operation is also conducted on the scribbles s_i and the prediction maps of the student network. The training procedure can be divided into two phases as well, i.e. training the ancillary stage and training the primary stage. First, to train the model in the ancillary stage, we use scribble annotation by minimizing the saliency loss L_{sal} and the output of the teacher network by consistency loss L_{con} for constraining the model. After the backpropagation of loss L_{sal} and loss L_{con} , the ancillary teacher network will be updated by the EMA. Based on the ancillary stage, we can generate multiple pseudo labels for the primary stage. Second, to train the network in the primary stage, we still adopt the scribble annotation, the output of the teacher network and multiple pseudo labels for supervision. In order to realize the information interaction between students, the output of the ancillary student network is also fed into the primary student network as pseudo labels, which achieves the student-student model proposed in this work.

3.2 Seminar learning framework

3.2.1 Mean teacher-based scribble network

For the ancillary stage, the teacher model and student model have identical architectures, as shown in Fig. 3, where we develop a cross attention guided network (CAGNet) for saliency detection. We choose ResNet-50 [18] as the backbone. Given an input image, we can

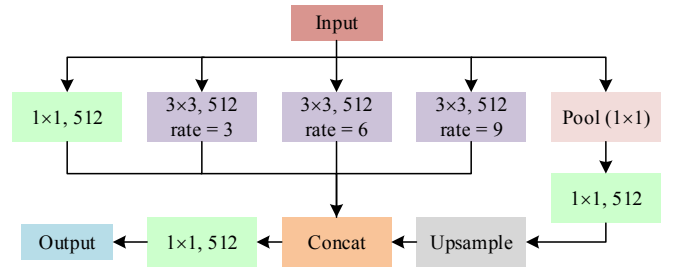


Fig. 4 Schematic illustration of the ASPP.

obtain features at different levels, denoted as $\{f_i | i = 1, \dots, 5\}$.

Using the U-shape architecture, there exists an issue that the high-level features will be gradually diluted through the top-down pathways. Therefore, we put the atrous spatial pyramid pooling (ASPP) module at the end of Conv5, as shown in Fig. 4. This is used to capture multi-scale context information using atrous convolutions and thus strengthen high-level information in the global view where we can obtain feature g_5 . For the ASPP module, the features extracted for each sampling rate are further processed in separate branches and fused to generate the final result. Specifically, there are five parallels, i.e. a 1×1 pointwise convolution, three 3×3 atrous convolutions with atrous rates of 3, 6 and 9, respectively, and an image pooling layer. All of these five parallel layers produce 512 channel feature maps. The image pooling layer consists of a global averaging pooling and a pointwise convolution that converts the number of output channels to 512. After that, these features concatenated together and followed by a pointwise convolution with 512 channels to obtain the output.

Motivated by the fact that neurons in the human brain connect in multiple directions, forming a gigantic network [2]. Therefore, we propose a series of CAF modules to integrate multi-level information and avoid undermining the effectiveness of information propaga-

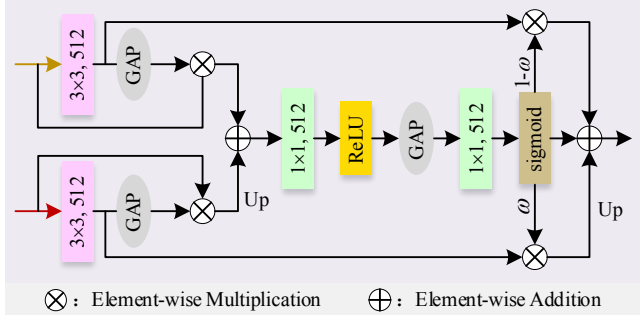


Fig. 5 Schematic illustration of the proposed cross attention fusion (CAF) module.

tion, as illustrated in Fig. 5. Different from SE module [22], our CAF module considers two different features at the same time, and integrates different levels of features using the thought of attention mechanism. The channel adjustment is employed with feature f_i and formulates the feature $\{f_i^{ca} | i = 1, \dots, 5\}$, which is used as an input to CAF. More specifically, to learn the important information of each input feature, a 3×3 convolution layer is applied and produce 512 channels feature map, respectively. Then a global average operation are convolved with feature g_5 and f_5^{ca} , respectively. Based on the weights β_1 and β_2 , we generate more useful and discriminative features by combining them, which can be written as:

$$F = \underbrace{\Phi(\text{conv}_{3 \times 3}(f_i^{ca}))}_{\beta_1} \cdot f_i^{ca} + \text{Up}(\underbrace{\Phi(\text{conv}_{3 \times 3}(g_5))}_{\beta_2} \cdot g_5), \quad (1)$$

where $\text{conv}_{3 \times 3}(\cdot)$ denotes the 3×3 convolution operation, Φ is the global average pooling operation and Up is the upsample operation. Then feature F is applied as an 1×1 convolution layer (with BN and ReLU) with 512 channels and a global average pooling operation. After that, an 1×1 convolution layer is used for forming a weight ω , then it will multiply with the former feature and obtained attention map. Finally, we combine them by pixel-wise summation for generating the final results. Similarly, the following CAFs are performed in the same manner. The process is formally described as follows:

$$\begin{aligned} S_1 &= \omega \cdot \text{Up}(\text{conv}_{3 \times 3}(f_g)) + (1 - \omega) \cdot \text{conv}_{3 \times 3}(f_5) \\ S_i &= \omega \cdot \text{Up}(\text{conv}_{3 \times 3}(S_{i-1})) \\ &\quad + (1 - \omega) \cdot \text{conv}_{3 \times 3}(f_{6-i}) \quad i = 2, 3, 4 \end{aligned} \quad (2)$$

where $\text{conv}_{1 \times 1}(\cdot)$ denotes the 1×1 convolution operation.

It is worth mentioning that the teacher model and the student model have the same architecture, which are both randomly initialized with the same random seed. The teacher network is updated by the EMA strategy of the student network such that:

$$\vartheta_j = \alpha \vartheta_{j-1} + (1 - \alpha) \theta_j, \quad (3)$$

where θ_j is the student network parameter at the j th training iteration while ϑ_{j-1} is the parameters of the teacher network at the $(j - 1)$ th training iteration in the ancillary stage, respectively, and $\alpha \in (0, 1)$ is a smoothing coefficient hyperparameter to control the updating rate. According to [54, 7], hyperparameter α should be specified. It is reasonable for it to be set to a value close to 1 and therefore, we set $\alpha = 0.999$ in our experiments.

For the second training stage, i.e. primary stage, we utilize a similar approach with the ancillary stage. The model still adopts the teacher-student network whose architecture is invariant, since it can be used to expand the learning range of the network. Like of the ancillary stage, the teacher network of the primary stage is able to prevent the student network from overfitting to scribble annotations. Note that we also propose the student-student model in this part to connect the ancillary student network and the primary student network using multiple pseudo labels, where we adopt different random seeds to generate different masks with significant diversity. The details can be found in Section 3.3.

3.2.2 Consistent transformation mechanism

Different input transformations (i.e. random rotation, flipping and scaling) are a prominent challenge for classification and segmentation problems in weakly supervised learning [29]. SOD requires the detection of an accurate target area, which belongs to one of the segmentation problems. As we all know, CNNs only distinguish the existence or not of an image for classification, it does not matter whether the image transformations are applied. However, transformations applied to the input image must be applied to the corresponding ground-truth in the SOD task, since the shape and size of target regions can vary significantly, which would influence the detection accuracy. As shown in Fig. 6, we can observe that prediction 1 and prediction 2 vary under the rotational change. Therefore, different input transformations would have an effect on SOD. Thus, we define a salient object detection as $f(\cdot)$ and a transformation as $T_i(\cdot)$. In ideal conditions, we can obtain:

$$f(T(x)) = T(f(x)), \quad (4)$$

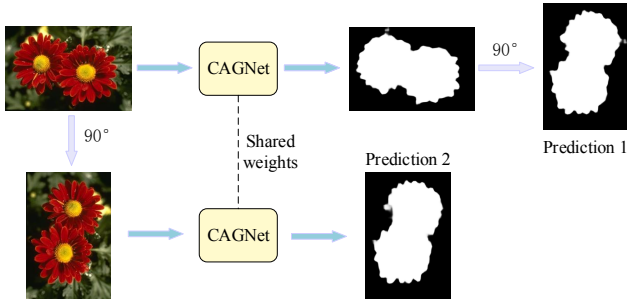


Fig. 6 Saliency maps under different conditions and we expect prediction 1 and prediction 2 to be equivariant.

that is, saliency maps are desired to be rotation equivariant. In view of these observations, we consider incorporating the consistent transformation mechanism to our framework to make prediction 1 and prediction 2 to be consistent.

As shown in Fig. 2, both teacher networks applied the consistent transformation operation $T_i(\cdot)$ to the input image, i.e. random rotation, random flipping and random scaling. For the student network in the ancillary and primary stage, the consistent transformation operation $T_i(\cdot)$ is applied to the predicted map. In this work, four rotation operations are adopted by using angles of $\gamma \times 90^\circ$, where $\gamma \in \{0, 1, 2, 3\}$, scaling operations use scaling ratios between 0.8 to 1.2, along with random flipping. Further, input perturbation plays an important role in the success of the self-fusion method. Hence, we utilize Gaussian noise in both teacher networks to strengthen the generalization capability. Note that our approach is different from traditional data augmentation, in other words, our method utilized the data without requiring its label by minimizing the prediction difference in the teacher-student models under various transformations.

3.3 Loss function

In this section, we will introduce the loss function of this model. It mainly consists of two parts, i.e. the loss function of ancillary stage and the loss function of primary stage.

3.3.1 Ancillary stage

In the ancillary stage, we use the teacher-student model for training. First, we train the student network using scribbles, which is defined as follows:

$$L_{sal} = L_{pce} + L_{lsc}, \quad (5)$$

where

$$L_{pce} = \sum_{i \in S} -\tilde{s}_i \log y_i^s + (1 - \tilde{s}_i) \log(1 - y_i^s), \quad (6)$$

and

$$L_{lsc} = \sum_i \sum_{j \in G_i} F(i, j) D(i, j), \quad (7)$$

where L_{pce} and L_{lsc} denote partial cross entropy loss [53] and local saliency coherence (LSC) loss [70], respectively, $\tilde{s}_i$ is the scribble annotations, y_i^s is the predicted values, S is the labeled pixel set, $F(i, j)$ is Gaussian kernels and $D(i, j)$ is L1 distance of pixels i and j . It is noted that L_{lsc} and L_{pce} are able to enforce the model to acquire enlarged salient regions with sparse labels.

Since the images undergo transformation operations before entering the teacher network, then the predictions of student network also undergo the same operations. To make them consistent, we need to compute consistency loss. In view of the above consideration, we adopt the mean square error (MSE) [69] to calculate the discrepancy between the teacher network and the student network, which is able to get the support of the teacher network. It is described as follows:

$$L_{con} = -\frac{1}{\Omega} \sum_{i \in \Omega} \|f_\theta(x_i) - f_\vartheta(x_i)\|^2, \quad (8)$$

where Ω is the total number of pixels in each image, $f_\theta(\cdot)$ is the softmax prediction of the student network, $f_\vartheta(\cdot)$ is the softmax prediction of the teacher network. Finally, the general optimization objective of the ancillary student network is able to be formulated as follows:

$$L_{anx} = \gamma L_{sal} + \gamma_1 L_{con}, \quad (9)$$

where λ and λ_1 are the hyper-parameters, L_{sal} and L_{pce} are a saliency prediction loss and a pixel consistency loss, respectively. After the backpropagation of loss L_{sal} and loss L_{con} , the ancillary teacher network will be updated by EMA.

3.3.2 Primary stage

As for the primary stage, we adopt the teacher-student and student-student model for joint training, where the student-student model requires an intermediate bridge. Early work [73] adopts unsupervised methods to generate pseudo labels for training the network. However, unsupervised methods would inevitably produce unreliable predictions owing to the fact there is no ground

Algorithm 1 The pseudocode of SLF-CT.

Input: training data $I = \{(x_i, s_i, p_i)\}_{i=1}^M$ $\triangleright$ original image $x_i \in X$, scribble annotation $s_i \in S$ and pseudo mask $p_i \in P$ from the last student model

Output: parameters: θ, ϑ $\triangleright$ the student model with parameters θ , the teacher model with parameters ϑ

```

1: for each epoch  $t$  in  $[1, MaxEpochs]$  do
2:   for each batch  $N$  do
3:     original image  $x_i$ , scribble annotation  $s_i$ , pseudo mask  $p_i = \text{Dataloader}(I)$   $\triangleright$  data loading
4:      $\tilde{x}_i \sim T_i(x_i)$ ,  $\tilde{s}_i \sim T_i(s_i)$ ,  $\tilde{p}_i \sim T_i(p_i)$   $\triangleright$  random transformation operation
5:      $\tilde{y}_i^t \sim f_{\vartheta}(\tilde{x}_i)$   $\triangleright$  prediction of the teacher module  $f_{\vartheta}(\cdot)$ 
6:      $y_i^s \sim f_{\theta}(x_i)$   $\triangleright$  prediction of the student model  $f_{\theta}(\cdot)$ 
7:      $\tilde{y}_i^s \sim T_i(y_i^s)$   $\triangleright$  transform the output of the student model
8:      $loss \leftarrow \text{Eq. (10)}$   $\triangleright$  calculate loss function
9:     Update  $\theta$  via SGD optimizer  $\triangleright$  update the parameters  $\theta$  of the student model
10:    Update  $\vartheta$  via Eq. (3)  $\triangleright$  update the parameters  $\vartheta$  of the teacher model
11:   end for
12: end for

```

truth. Different from [73], we generate multiple pseudo labels by the ancillary student network with different initial seeds to the primary student network, which is used to produce different masks. It is essentially a weakly supervised algorithm. Such a discrimination model with less confirmation bias can correctly guide the student network training in the primary stage. The general optimization objective of the primary student network can be described as follows:

$$L_{dom} = \gamma L_{sal} + \gamma_1 L_{con} + \gamma_2 L_{pse}, \quad (10)$$

where γ_2 is the hyper-parameter, other parameters' meaning are the same as Eq. (9). L_{pse} is formulated by the binary cross entropy function [77] because the pseudo labels are applied to each pixel, which is defined as:

$$L_{pse} = - \sum_{(x,y)} [p_i(x,y) \log(y_i^s(x,y)) + (1 - p_i(x,y)) \log(1 - y_i^s(x,y))], \quad (11)$$

where $p_i(x,y)$ and $y_i^s(x,y)$ denote the pseudo ground truth masks and the predicted saliency maps, respectively. The primary teacher network is also updated by EMA.

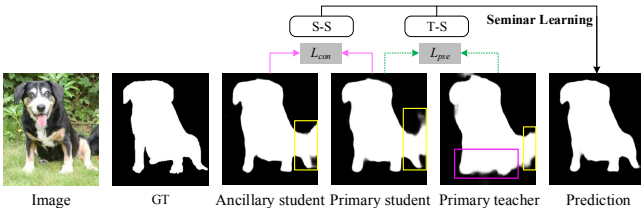


Fig. 7 Visualization of the strategy in seminar learning. Here, 'S-S' means student-student module and 'T-S' means teacher-student module.

Discussion The main steps of the proposed SLF-CT are summarized in Algorithm 1. We further demonstrate some examples of each network in the training of weakly supervised SOD in Fig. 7. It is worth to mention that the saliency map is more highlighted after seminar learning, endorsing our approach to use seminar learning. Despite the ancillary student learning the approximate location, there are still details to be optimized, such as the dog's leg in the yellow box. With respect to the primary teacher, it occupies a wider areas of the dog compared with the student which proves the better generalization capacity, such as the dog with pink box. In general, benefiting from the communication between the students and the teachers, as well as other students, our SLF-CT achieves significant and superior results.

4 Experiments

4.1 Implementation details

The proposed approach is implemented on the Pytorch platform using an RTX3090(24G) GPU. The batch size is set as 32. The student network is optimized using stochastic gradient descent (SGD) and the teacher network is optimized by EMA of the student network, where the weight decay is set to 5e-4, the momentum is set to 0.9 and the initial learning rate is 1e-5. We set the hyperparameters of the loss weights as $\gamma = 5$, $\gamma_1 = \gamma_2 = 1$. We set the EMA decay parameter $\alpha = 0.999$ for each teacher-student module. We resize each image to 320×320 and feed into the network, then the final saliency maps is restored to its original size. Note that both the ancillary stage and the primary stage, all the network adopts the same hyperparameters. Additionally, the characteristics of each dataset are summarized in Table 1.

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

Name	Stage	Size	Description
S-DUTS [74]	Train	10553	Relabeled salient object detection dataset DUTS-TR with scribbles.
ECSSD [67]	Test	1000	Dataset includes semantically meaningful and complex structures.
DUT-OMRON [68]	Test	5168	Dataset with high quality and challenging images having one or more salient objects with complex background scenes.
PASCAL-S [13]	Test	850	A dataset selected from the PASCAL VOC 2010 segmentation, which contains 20 object categories and complex scenes.
HKU-IS [25]	Test	4447	Contains multiple salient objects with overlapping objects touching the image boundary or with low color contrast.
DUTS-TE [55]	Test	5019	Selected from the largest salient object detection benchmark dataset DUTS, which contains complex scens in different scales.

4.2 Evaluation metrics

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

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

$$F_\beta = \frac{(1 + \beta^2) \times Precision \times Recall}{\beta^2 \times Precision + Recall}, \quad (12)$$

where *Precision* and *Recall* are given by thresholding the predicted saliency map, and β^2 is set to 0.3 in accordance with [8]. Then the obtained pairs (threshold, F_β) is employed to plot the **F-measure curve**.

$$Precision = \frac{TP}{TP + FP}, \quad Recall = \frac{TP}{TP + FN}, \quad (13)$$

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

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

$$MAE = \frac{1}{W \times H} \sum_{x=1}^W \sum_{y=1}^H |S(x, y) - G(x, y)|, \quad (14)$$

where W and H represent width and height of the saliency maps respectively.

Enhanced-alignment measure E_ϕ [15] is applied to evaluate both local and global similarity between the predicted map and the ground-truth:

$$E_\phi = \frac{1}{W \times H} \sum_{x=1}^W \sum_{y=1}^H F_\phi(x, y), \quad (15)$$

where F_ϕ denotes the enhanced alignment matrix.

Structure measure S_α [14] is utilized as the structure similarity of the predicted non-binary saliency map and the ground-truth, which is defined as follows:

$$S_\alpha = (1 - \alpha)S_r + \alpha S_o, \quad (16)$$

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

Furthermore, we evaluate the efficiency and flexibility of various methods, including the number of model parameters (#PARM), the number of floating point operations (FLOPs), the model size and the speed (Frames Per Second, FPS). The number of FLOPs measures the computational cost of a model, and a smaller number of FLOPs means lower energy consumption. Note that #PARM, the model size and FLOPs are measured in millions (M), MBbyte (MB) and giga (G), respectively.

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

We compare our approach with twenty five methods, including fourteen fully supervised methods (Amulet [77], UCF [78], NLDF [41], RAS [8], PAGR [79], BMPM [76], DSS [20], EGNet [80], CPD [63], MINet [46], VST [35], CII [34], DNA [38] and ICON [82]), two unsupervised methods (SVF [73] and C2S [28]), nine weakly supervised methods (WSS [55], ASMO [24], MWS [71], WSSA [74], MFNet [47] and SBBs [39]).

Quantitative comparison The detailed F-measure, MAE, E-measure and S-measure values are illustrated in Table 2 and Table 3 on five benchmarking datasets. For weakly supervised SOD methods, the top three best entries are marked in red, blue, and green, respectively. We can see that SLF-CT is able to consistently outperform other state-of-the-art unsupervised and

Table 2 Comparison with other state-of-the-art approaches on ECSSD and DUT-OMRON datasets. 'F' means fully-supervised learning, 'U' means unsupervised learning and 'W' means weakly-supervised learning. $\uparrow$ & $\downarrow$ denote larger and smaller is better, respectively.

Methods	Year	Sup.	ECSSD				DUT-OMRON			
			$F_\beta \uparrow$	MAE $\downarrow$	$E_\phi \uparrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	MAE $\downarrow$	$E_\phi \uparrow$	S_α
UCF [78]	2017	F	0.844	0.069	0.892	0.883	0.621	0.120	0.765	0.760
Amulet [77]	2017	F	0.868	0.059	0.901	0.894	0.647	0.098	0.779	0.781
NLDF [41]	2017	F	0.878	0.063	0.910	0.875	0.684	0.080	0.816	0.770
RAS [8]	2018	F	0.889	0.056	0.914	0.893	0.713	0.062	0.846	0.814
PAGR [79]	2018	F	0.894	0.061	0.914	0.889	0.711	0.071	0.842	0.775
BMPM [76]	2018	F	0.868	0.045	0.914	0.911	0.692	0.064	0.837	0.809
DSS [20]	2019	F	0.904	0.052	0.912	0.882	0.740	0.063	0.842	0.790
EGNet [80]	2019	F	0.920	0.037	0.927	0.925	0.755	0.053	0.868	0.841
CPD [63]	2019	F	0.917	0.037	0.925	0.918	0.747	0.056	0.866	0.825
MINet [46]	2020	F	0.924	0.033	0.927	0.925	0.755	0.056	0.865	0.833
VST [35]	2021	F	0.920	0.033	0.918	0.932	0.756	0.058	0.861	0.850
CII [34]	2021	F	0.929	0.033	0.926	0.926	0.768	0.054	0.872	0.839
DNA [38]	2022	F	0.920	0.035	0.928	0.918	0.743	0.056	0.857	0.826
ICON [82]	2022	F	0.928	0.032	0.927	0.927	0.772	0.057	0.870	0.844
SVF [73]	2017	U	0.809	0.088	0.875	0.832	0.608	0.108	0.768	0.747
C2S [28]	2018	U	0.853	0.059	0.906	0.882	0.664	0.079	0.817	0.780
WSS [55]	2017	W	0.823	0.104	0.869	0.811	0.603	0.109	0.768	0.725
ASMO [24]	2018	W	0.798	0.110	0.853	0.802	0.622	0.101	0.776	0.752
MSW [71]	2019	W	0.840	0.096	0.884	0.827	0.609	0.109	0.763	0.756
WSSA [74]	2020	W	0.870	0.059	0.901	0.865	0.703	0.068	0.840	0.785
MFNet [47]	2021	W	0.844	0.084	0.877	0.837	0.621	0.098	0.783	0.726
SBBs [39]	2021	W	0.855	0.072	0.894	0.851	0.695	0.074	0.835	0.776
OSG [58]	2021	W	0.893	0.058	0.900	0.866	0.728	0.068	0.847	0.777
NSAL [48]	2022	W	0.856	0.078	0.883	0.834	0.648	0.088	0.801	0.745
SHNet [43]	2022	W	0.758	0.126	0.817	0.750	0.641	0.094	0.763	0.704
Ours	-	W	0.878	0.056	0.908	0.864	0.708	0.066	0.844	0.778

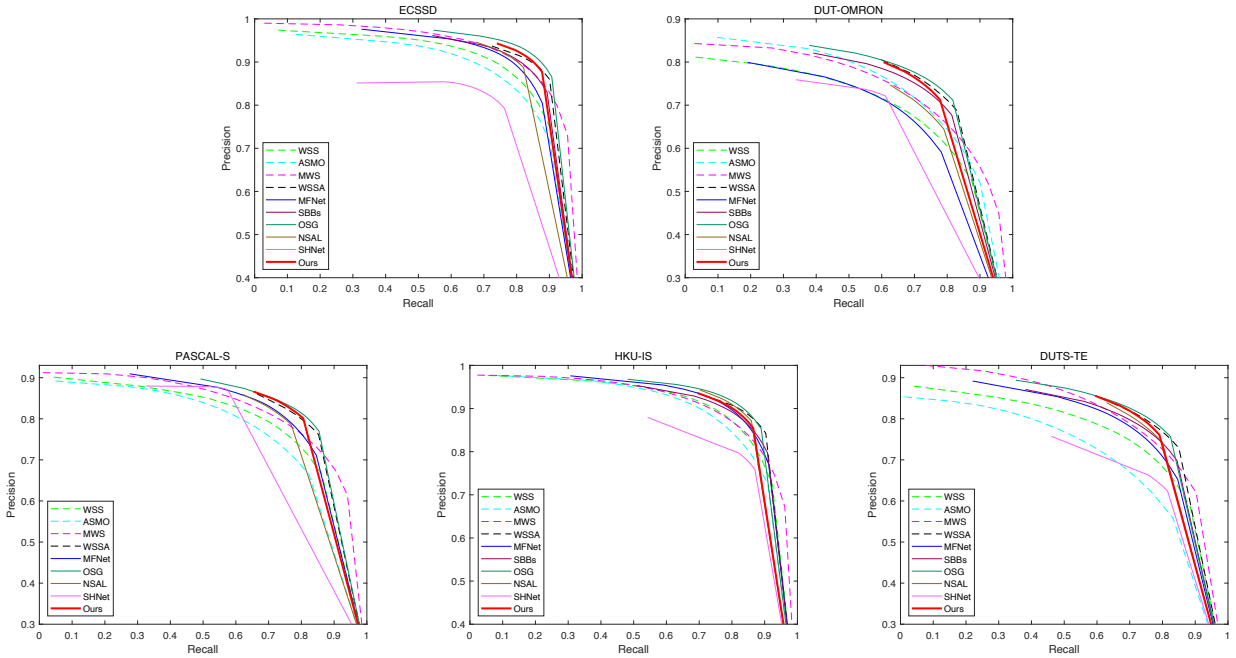


Fig. 8 PR curves of the proposed approach with other state-of-the-art weakly supervised methods using five datasets.

weakly supervised approaches in most cases, even superior to some fully supervised methods, such as Amulet,

UCF and NLDF, which demonstrates its effectiveness. In particular, our proposed approach outperformed the

Table 3 Comparison with other state-of-the-art approaches on PASCAL-S, HKU-IS and DUTS-TE datasets. 'F' means fully-supervised learning, 'U' means unsupervised learning and 'W' means weakly-supervised learning. $\uparrow$ & $\downarrow$ denote larger and smaller is better, respectively. "-" means the authors did not provided the saliency maps, thus reporting the evaluation indices of this method is not possible.

Methods	Year	Sup.	PASCAL-S				HKU-IS				DUTS-TE			
			$F_\beta \uparrow$	MAE $\downarrow$	$E_\phi \uparrow$	S_α	$F_\beta \uparrow$	MAE $\downarrow$	$E_\phi \uparrow$	S_α	$F_\beta \uparrow$	MAE $\downarrow$	$E_\phi \uparrow$	S_α
UCF [78]	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
Amulet [77]	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
NLDF [41]	2017	F	0.769	0.098	0.839	0.805	0.874	0.048	0.929	0.887	-	-	-	-
RAS [8]	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 [79]	2018	F	0.798	0.089	0.853	0.822	0.886	0.048	0.939	0.887	0.786	0.054	0.880	0.838
BMPM [76]	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 [20]	2019	F	0.801	0.093	0.847	0.798	0.902	0.040	0.934	0.878	-	-	-	-
EGNet [80]	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 [63]	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 [46]	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 [35]	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
CH [34]	2021	F	0.841	0.062	0.863	0.865	0.915	0.029	0.955	0.920	0.835	0.037	0.900	0.888
DNA [38]	2022	F	0.815	0.074	0.856	0.837	0.903	0.028	0.950	0.913	0.813	0.039	0.892	0.872
ICON [82]	2022	F	0.833	0.064	0.861	0.861	0.910	0.029	0.952	0.920	0.838	0.037	0.902	0.889
SVF [73]	2017	U	0.695	0.131	0.789	0.758	-	-	-	-	-	-	-	-
C2S [28]	2018	U	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 [55]	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 [24]	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
MSW [71]	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 [74]	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 [47]	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 [39]	2021	W	-	-	-	-	0.843	0.056	0.920	0.854	0.722	0.073	0.851	0.789
OSG [58]	2021	W	0.792	0.090	0.830	0.798	0.878	0.050	0.926	0.848	0.763	0.063	0.856	0.786
NSAL [48]	2022	W	0.756	0.110	0.823	0.767	0.864	0.051	0.923	0.854	0.730	0.073	0.849	0.782
SHNet [43]	2022	W	0.699	0.147	0.732	0.697	0.777	0.073	0.883	0.806	0.637	0.094	0.797	0.733
Ours	-	W	0.789	0.087	0.841	0.794	0.857	0.050	0.923	0.851	0.750	0.060	0.865	0.793

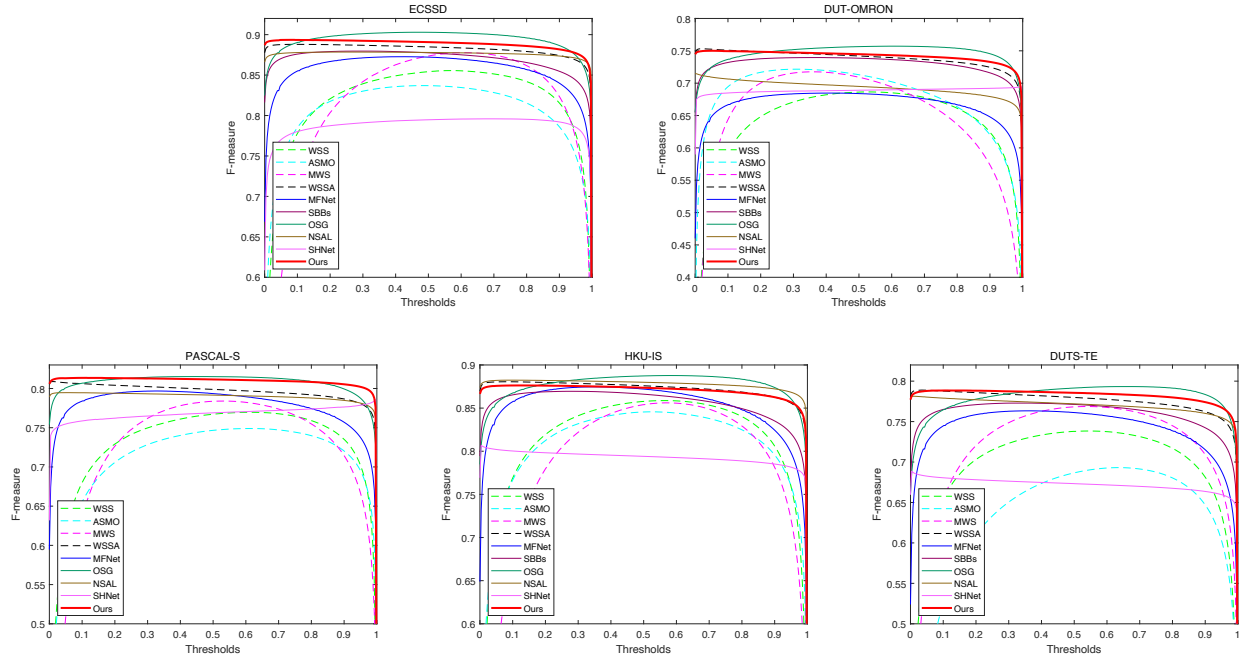


Fig. 9 F-measure curves of the proposed approach with other state-of-the-art weakly supervised methods using five datasets.

latest SBBs method with improvements of up to 2.7% (22.2%), 1.8% (10.8%), 3.9% (17.8%) when using the F-measure (MAE) score on ECSSD, DUT-OMRON and DUTS-TE dataset, respectively. Although our method is not the best in the HKU-IS dataset, it is second only to WSSA. As expected, there are some gaps compared with the fully supervised algorithm in terms of performance even though it is reasonable and logical, for example, CII and ICON. We can observe that SHNet is poor compared with other methods, it may be that SHNet focuses on optical remote sensing images not natural scene images. Generally speaking, our approach is superior to other counterparts across all datasets using these evaluation metrics.

PR curves and F-measure curves with other weakly supervised state-of-the-art methods on the five benchmark datasets are shown in Fig. 8 and Fig. 9. It can be observed that SLF-CT performs favorably against the other existing models in most cases. Note that SLF-CT achieves comparable performance with respect to the latest weakly supervised SOD model NSAL.

For additional comparison, we further analyze the overall difference between our algorithm and other state-of-the-art fully supervised methods. We compare the average F-measure on five datasets between SLF-CT and other fully supervised methods, as illustrated in Fig. 10. It can be easily seen that we do not achieve overall best performance, but we do gain comparative or better performance in many cases than fully supervised techniques. In particular, our SLF-CT achieves 92.8% accuracy of fully supervised methods at least. It is worth noting that our approach requires much less effort in terms of data annotation.

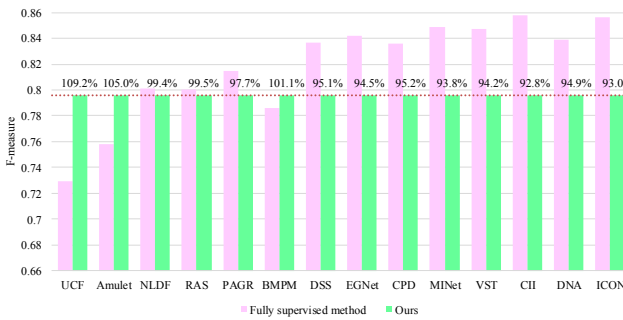


Fig. 10 Average F-measure comparison of our method with 11 fully supervised methods. The pink part represents the performance of each fully supervised methods and the green one indicates ours.

Qualitative comparison We present some examples of saliency maps from our proposed method and other state-of-the-art methods using some challenging cases in Fig. 11, including simple scenes (1st row), large

objects (2nd and 3rd rows), cluttered backgrounds (4th and 9th rows), low contrast (5th and 6th rows) multiple targets (8th row) and complex scenes (7th and 10th rows). Specifically, the 1st row shows an image with a lady leaning against the wall in an image, almost all methods incorrectly generate satisfactory results except our method. In the 2nd and 3rd rows, most methods can not locate the entire salient regions although there is a big object. Compared with the 1st to 3rd rows, a pot of beans presented in the 4th row is more difficult to segment due to cluttered backgrounds. As we can see, our SLF-CT is able to accurately find the salient object with fewer false salient pixels detected. Nevertheless, fully supervised methods are hardly to do it. There is a squirrel that its color and background are very similar in the 5th row. We are able to see that our proposed method can effectively suppress the background noise and accurately detect the salient target. Similarly, the 6th row corresponds to a crane by the river, where there is low contrast between the target and image background. It is easy to be observed that our model achieves good performance. Different from the previous examples, there are two women as salient objects in the 8th row. The proposed method can generate reliable saliency maps. Furthermore, as shown in the 7th row, the proposed method can also identify accurate locations when there are more background objects.

Generally speaking, compared with other weakly supervised methods and unsupervised methods, our SLF-CT performs better with respect to salient object segmentation and localization. It is worth mentioning that our proposed approach is also able to generate better prediction than fully supervised methods in some cases.

Table 4 Comparison results of the proposed method with other methods.

Methods	Size	#Param (M)	Size (MB)	FLOPs (G)	FPS
EGNet [80]	320 × 320	111.66	447.1	244.14	20
CPD [63]	320 × 320	47.85	192.0	14.66	38
MINet [46]	320 × 320	162.38	650.0	87.03	38
VST [35]	320 × 320	44.03	178.4	23.16	29
WSSA [74]	320 × 320	15.53	64.1	64.45	76
MFNet [47]	320 × 320	16.35	-	16.01	-
SHNet [43]	356 × 356	20.71	88.0	110.64	25
Ours	320 × 320	93.93	376.4	13.59	62

4.4 Network complexity and inference speed

For understanding the network more comprehensively of this paper, we also evaluate the efficiency and flex-

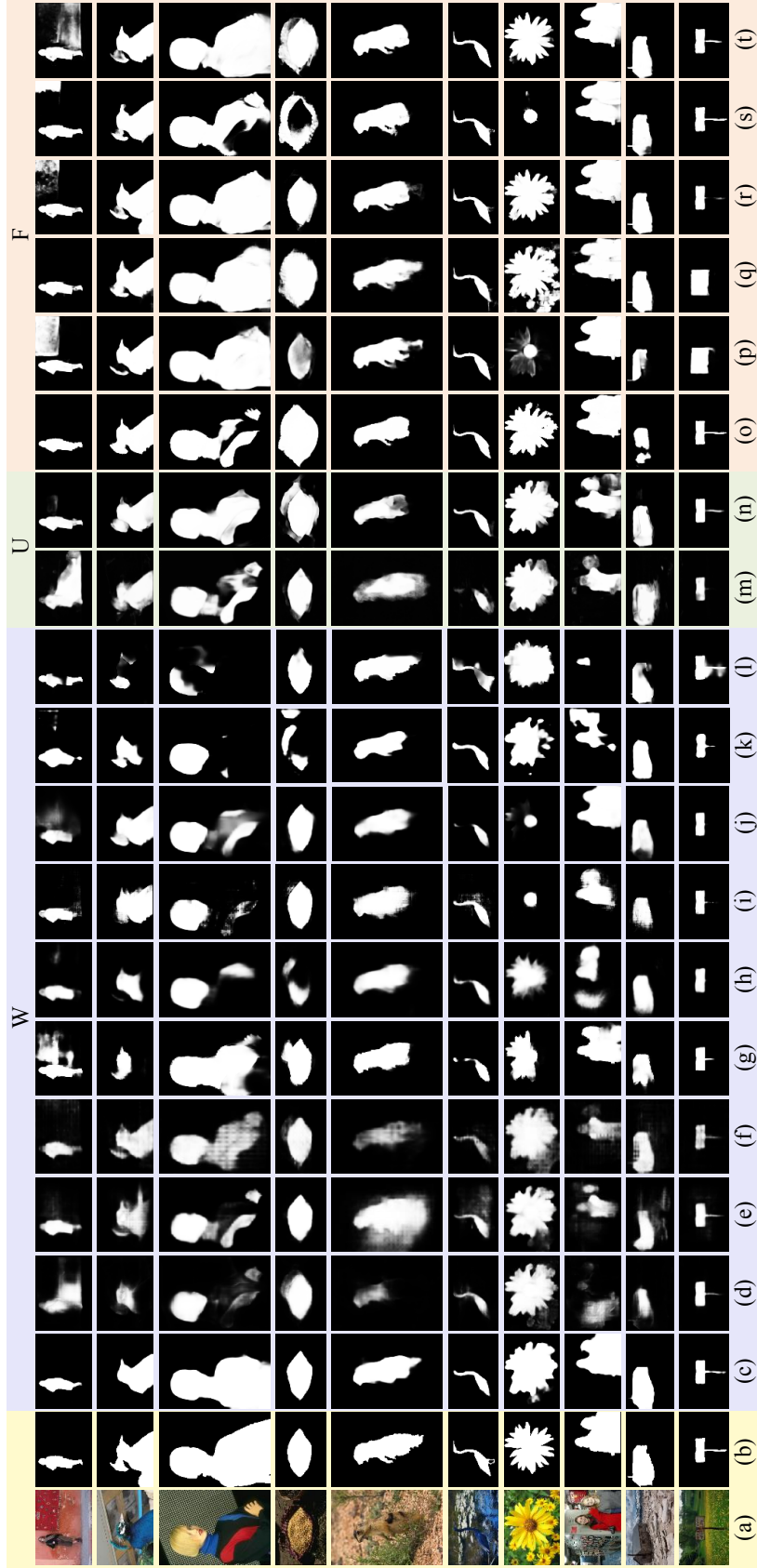


Fig. 11 Visual comparison between our method and other state-of-the-art methods. (a) Image; (b) GT (ground-truth); (c) Ours; (d) WSS; (e) ASMO; (f) MSW; (g) WSSA; (h) MFNet; (i) SBBs; (j) OSG; (k) NSAL; (l) SHNet; (m) SVF; (n) C2S; (o) DSS; (p) CPD; (q) VST; (r) CII; (s) DNA; (t) ICON. Here, 'W', 'U' and 'F' have the same meaning as above.

Table 5 Ablation studies of different EMA decay weight α on ECSSD and DUT-OMRON datasets. The best results are highlighted in **bold**.

	ECSSD				DUT-OMRON			
	$F_\beta \uparrow$	MAE $\downarrow$	$E_\phi \uparrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	MAE $\downarrow$	$E_\phi \uparrow$	$S_\alpha \uparrow$
$\alpha = 0.9$	0.578	0.202	0.678	0.635	0.412	0.234	0.590	0.560
$\alpha = 0.99$	0.761	0.097	0.825	0.790	0.603	0.113	0.753	0.724
Ours	0.878	0.056	0.908	0.864	0.708	0.066	0.844	0.778

Table 6 Ablation studies of different EMA decay weight α on PASCAL-S, HKU-IS and DUTS-TE datasets. The best results are highlighted in **bold**.

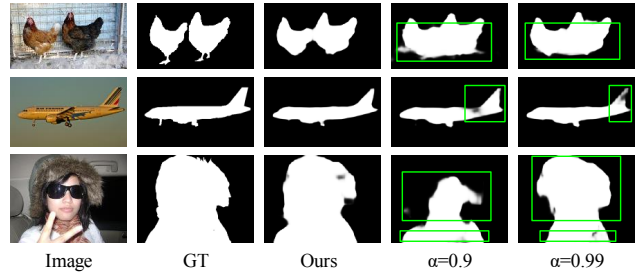
	PASCAL-S				HKU-IS				DUTS-TE			
	$F_\beta \uparrow$	MAE $\downarrow$	$E_\phi \uparrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	MAE $\downarrow$	$E_\phi \uparrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	MAE $\downarrow$	$E_\phi \uparrow$	$S_\alpha \uparrow$
$\alpha = 0.9$	0.513	0.229	0.625	0.586	0.554	0.181	0.707	0.642	0.413	0.217	0.606	0.566
$\alpha = 0.99$	0.700	0.125	0.774	0.748	0.727	0.088	0.837	0.779	0.605	0.110	0.758	0.723
Ours	0.789	0.087	0.841	0.794	0.857	0.050	0.923	0.851	0.750	0.060	0.865	0.793

ibility of different methods. Since not all of methods providing source code, we only compare the results of some methods. Table 4 shows the comparison results of various methods, they all run a device with RTX3090 GPU. The average run time for each method is 5000 executions on the 320×320 size images (except for SHNet where the image size is 356×356). Other evaluation metrics also adopt the same resolution. From these results, we can see that our model has 376.4 MB that leads to it requiring the amount of space, this is the limitation of our method. Nevertheless, our method has the smallest FLOPs, which means that the lowest energy consumption can be used. As for speed, although the proposed method is not the fastest, it still exceeds almost all methods and achieves real-time, especially fully supervised methods. In general, our method has certain advantages to some extent.

4.5 Ablation studies

In this section, we perform a series of tests using five datasets to assess the effectiveness of our proposed algorithm. All the ablation studies follow the same implementation setup.

Effect on hyper-parameter setting. First, we conduct various experiments to verify the effectiveness of different EMA weights α using five datasets. From Table 5 and Table 6, we can conclude that the performance is better as α is larger. Compared with $\alpha = 0.99$, F-measure has been improved by 12.7% using the PASCAL-S dataset and E-measure has been improved from 0.678 to 0.908 in ECSSD dataset. Clearly, our proposed method obtains the best results in all datasets when setting $\alpha = 0.999$. Fig. 12 provides some visual examples to demonstrate the influence under different EMA decay weights α . It is clear to see that the main differences in each row, which have been marked using

**Fig. 12** Visual examples of different EMA decay weight α .

green boxes. For instance, other settings are difficult to detect the accurate location of two chicks, but our method can do it.

Second, to gain a better understanding of the proposed approach, Table 7 and Table 8 demonstrate the ablation results of various loss function configurations. It can be seen that the model achieves excellent accuracy and prediction by setting $\gamma = 5, \gamma_1 = \gamma_2 = 1$. We also find that if the same weight is used ($\gamma = \gamma_1 = \gamma_2 = 1$), the performance of the model will not increase, which may be that L_{sal} is small weight and has weakened impact to the model. Specifically, the MAE of our model will decrease 98% when compared our method with $\gamma = \gamma_1 = \gamma_2 = 1$ on HKU-IS datasets, which indicates that the saliency map of our model is much closer to ground-truth. As for other settings, the overall effect of them is not the best. These results verify the rationality of our settings.

The number of pseudo labels. The number of the pseudo labels ($\tilde{N}$) determines the performance and generalization of the network. The experimental results are reported in Table 9 and Table 10. $\tilde{N} = 0$ denotes that the number of pseudo labels is zero, that is to say, the framework has no student-student model, hence its performance would be decreased. When applying a pseudo label, the F-measure of the network improves from 0.715 to 0.732 over DUTS-TE datasets. We also

Table 7 Performance comparison with different parameters of loss function on ECSSD and DUT-OMRON datasets. The best results are highlighted in **bold**.

γ	γ_1	γ_2	ECSSD				DUT-OMRON			
			$F_\beta \uparrow$	MAE $\downarrow$	$E_\phi \uparrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	MAE $\downarrow$	$E_\phi \uparrow$	$S_\alpha \uparrow$
1	1	1	0.757	0.103	0.833	0.795	0.546	0.140	0.705	0.690
3	1	1	0.860	0.062	0.899	0.856	0.677	0.076	0.825	0.766
5	1	1	0.878	0.056	0.908	0.864	0.708	0.066	0.844	0.778
7	1	1	0.854	0.065	0.886	0.852	0.668	0.089	0.804	0.761
5	0.5	1	0.846	0.079	0.874	0.816	0.681	0.073	0.828	0.742
5	3	1	0.831	0.072	0.883	0.845	0.634	0.093	0.791	0.747
5	1	0.5	0.860	0.069	0.901	0.865	0.668	0.074	0.819	0.783
5	1	3	0.861	0.062	0.895	0.856	0.685	0.072	0.832	0.772

Table 8 Performance comparison with different parameters of loss function on PASCAL-S, HKU-IS and DUTS-TE datasets.. The best results are highlighted in **bold**.

γ	γ_1	γ_2	PASCAL-S				HKU-IS				DUTS-TE			
			$F_\beta \uparrow$	MAE $\downarrow$	$E_\phi \uparrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	MAE $\downarrow$	$E_\phi \uparrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	MAE $\downarrow$	$E_\phi \uparrow$	$S_\alpha \uparrow$
1	1	1	0.676	0.136	0.760	0.736	0.714	0.099	0.838	0.779	0.559	0.129	0.721	0.700
3	1	1	0.769	0.092	0.829	0.791	0.836	0.055	0.915	0.844	0.716	0.069	0.846	0.782
5	1	1	0.789	0.087	0.841	0.794	0.857	0.050	0.923	0.851	0.750	0.060	0.865	0.793
7	1	1	0.754	0.104	0.804	0.774	0.833	0.058	0.908	0.842	0.703	0.077	0.827	0.773
5	0.5	1	0.742	0.110	0.802	0.738	0.820	0.069	0.879	0.797	0.715	0.073	0.839	0.748
5	3	1	0.734	0.109	0.805	0.771	0.800	0.067	0.896	0.829	0.679	0.080	0.820	0.772
5	1	0.5	0.767	0.096	0.831	0.799	0.837	0.061	0.915	0.855	0.700	0.070	0.832	0.796
5	1	3	0.759	0.095	0.821	0.782	0.837	0.054	0.914	0.844	0.717	0.069	0.846	0.781

Table 9 Performance with different numbers of pseudo labels on ECSSD and DUT-OMRON datasets. The best results are highlighted in **bold**.

Number ($\tilde{N}$)	ECSSD				DUT-OMRON			
	$F_\beta \uparrow$	MAE $\downarrow$	$E_\phi \uparrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	MAE $\downarrow$	$E_\phi \uparrow$	$S_\alpha \uparrow$
0	0.865	0.064	0.903	0.866	0.673	0.074	0.822	0.777
1	0.870	0.057	0.902	0.862	0.683	0.073	0.831	0.768
2	0.878	0.056	0.908	0.864	0.708	0.066	0.844	0.778
3	0.875	0.057	0.908	0.863	0.698	0.068	0.842	0.775

Table 10 Performance with different numbers of pseudo labels on PASCAL-S, HKU-IS and DUTS-TE datasets.. The best results are highlighted in **bold**.

Number ($\tilde{N}$)	PASCAL-S				HKU-IS				DUTS-TE			
	$F_\beta \uparrow$	MAE $\downarrow$	$E_\phi \uparrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	MAE $\downarrow$	$E_\phi \uparrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	MAE $\downarrow$	$E_\phi \uparrow$	$S_\alpha \uparrow$
0	0.775	0.090	0.833	0.803	0.845	0.056	0.920	0.856	0.715	0.066	0.841	0.797
1	0.783	0.086	0.838	0.796	0.850	0.051	0.921	0.849	0.732	0.064	0.855	0.787
2	0.789	0.087	0.841	0.794	0.857	0.050	0.923	0.851	0.750	0.060	0.865	0.793
3	0.782	0.089	0.842	0.793	0.852	0.051	0.923	0.849	0.743	0.061	0.862	0.793

observe that the model with $\tilde{N} = 0$ has good S-measure. Similarly, we utilize three pseudo labels for supervision to translate information between students, which illustrates that our proposed method is able to effectively improve the prediction of saliency maps. It concludes that more pseudo labels are not necessarily effective, because excessive pseudo labels will cause the model to be more weighted to the auxiliary stage and it will fail to achieve the role of optimization.

The number of student-student modules. Table 11 and Table 12 show the performance with different numbers of student-student models ($\hat{N}$). When $\hat{N} = 0$, it has the same meaning as the count of pseudo labels

being zero, hence, they have the same scores of evaluation metrics. Moreover, we find that the model with one student-student model with achieving the best results. When applying two or more student-student models, the F-measure and E-measure show little increase, and MAE does not decrease. That may be repeated using the student-student model to make the network rely on past results excessively and thus it is harder to obtain new knowledge. Fig. 13 presents some visual examples with different number of student-student models. And it can be seen that our SLF-CT obtains better results for segmentation, in particular, multiple objects. Note that for the 1st line, although our method does not de-

Table 11 Performance with different numbers of student-student model on ECSSD and DUT-OMRON datasets. The best results are highlighted in **bold**.

Number ($\hat{N}$)	ECSSD				DUT-OMRON			
	$F_\beta \uparrow$	MAE $\downarrow$	$E_\phi \uparrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	MAE $\downarrow$	$E_\phi \uparrow$	$S_\alpha \uparrow$
0	0.865	0.064	0.903	0.866	0.673	0.074	0.822	0.777
1	0.878	0.056	0.908	0.864	0.708	0.066	0.844	0.778
2	0.869	0.060	0.897	0.854	0.696	0.070	0.839	0.771

Table 12 Performance with different numbers of student-student model on PASCAL-S, HKU-IS and DUTS-TE datasets. The best results are highlighted in **bold**.

Number ($\hat{N}$)	PASCAL-S				HKU-IS				DUTS-TE			
	$F_\beta \uparrow$	MAE $\downarrow$	$E_\phi \uparrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	MAE $\downarrow$	$E_\phi \uparrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	MAE $\downarrow$	$E_\phi \uparrow$	$S_\alpha \uparrow$
0	0.775	0.090	0.833	0.803	0.845	0.056	0.920	0.856	0.715	0.066	0.841	0.797
1	0.789	0.087	0.841	0.794	0.857	0.050	0.923	0.851	0.750	0.060	0.865	0.793
2	0.778	0.088	0.836	0.787	0.846	0.054	0.913	0.836	0.739	0.064	0.857	0.783

Table 13 Performance with different types of transformations on ECSSD and DUT-OMRON datasets. The best results are highlighted in **bold**.

Transformation operations	ECSSD				DUT-OMRON			
	$F_\beta \uparrow$	MAE $\downarrow$	$E_\phi \uparrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	MAE $\downarrow$	$E_\phi \uparrow$	$S_\alpha \uparrow$
Baseline	0.844	0.065	0.891	0.840	0.626	0.090	0.795	0.737
Flipping	0.853	0.069	0.891	0.841	0.696	0.070	0.839	0.766
Scaling	0.867	0.059	0.902	0.857	0.691	0.069	0.840	0.772
Rotation	0.869	0.058	0.902	0.858	0.681	0.072	0.832	0.761
Ours	0.878	0.056	0.908	0.864	0.708	0.066	0.844	0.778

Table 14 Performance with different types of transformations on PASCAL-S, HKU-IS and DUTS-TE datasets. The best results are highlighted in **bold**.

Transformation operations	PASCAL-S				HKU-IS				DUTS-TE			
	$F_\beta \uparrow$	MAE $\downarrow$	$E_\phi \uparrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	MAE $\downarrow$	$E_\phi \uparrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	MAE $\downarrow$	$E_\phi \uparrow$	$S_\alpha \uparrow$
Baseline	0.734	0.101	0.812	0.769	0.807	0.062	0.902	0.830	0.682	0.075	0.826	0.768
Flipping	0.754	0.105	0.820	0.764	0.837	0.057	0.910	0.830	0.721	0.069	0.850	0.771
Scaling	0.780	0.089	0.837	0.791	0.843	0.053	0.919	0.843	0.726	0.064	0.854	0.783
Rotation	0.780	0.089	0.845	0.790	0.846	0.052	0.920	0.844	0.728	0.064	0.855	0.781
Ours	0.789	0.087	0.841	0.794	0.857	0.050	0.923	0.851	0.750	0.060	0.865	0.793

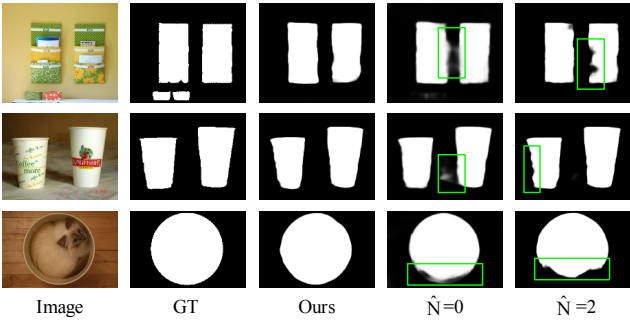


Fig. 13 Visual examples of different number of student-student models.

test the object at the bottom, our method is still the best in comparison.

Effect on different types of transformations. To introduce the contribution of the proposed consistent transformation mechanism, we train the network using different types of transformation operations, as illustrated in Table 13 and Table 14. "Baseline" refer-

s to the network without consistent transformations. "Flipping" refers to the baseline with the random flipping operation. Similarly, "Scaling" and "Rotation" refer to the baseline with the random scaling and rotation operation, respectively. First, the network with transformation operation performs better than the baseline. Specifically, the baseline with flipping transformation can effectively increase F-measure (from 0.626 to 0.696 on DUT-OMRON dataset) than baseline. The random rotation operation also provides an improved performance when compared to the baseline on various datasets. Compared with the baseline, using random rotation operation can obtain an average gain of 3.0% for F-measure, 10.8% for MAE, 1.2% for E-measure and 2.1% for S-measure on ECSSD dataset, respectively, which reveals that between the latter and the ground-truth has more local and global similarity. Second, we can see that scaling and rotation operation are superior to flipping. For example, the random rotation operation can effectively reduce MAE (from 0.105 to 0.089 on

Table 15 Ablation studies for different moduels on ECSSD and DUT-OMRON datasets. The best results are highlighted in **bold**.

ASPP	CAF	ECSSD				DUT-OMRON			
		$F_\beta \uparrow$	MAE $\downarrow$	$E_\phi \uparrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	MAE $\downarrow$	$E_\phi \uparrow$	$S_\alpha \uparrow$
✓		0.866	0.057	0.905	0.865	0.676	0.074	0.824	0.772
	✓	0.870	0.059	0.903	0.861	0.680	0.069	0.839	0.766
✓	✓	0.878	0.056	0.908	0.864	0.708	0.066	0.844	0.778

Table 16 Ablation studies for different modules on PASCAL-S, HKU-IS and DUTS-TE datasets. The best results are highlighted in **bold**.

ASPP	CAF	PASCAL-S				HKU-IS				DUTS-TE			
		$F_\beta \uparrow$	MAE $\downarrow$	$E_\phi \uparrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	MAE $\downarrow$	$E_\phi \uparrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	MAE $\downarrow$	$E_\phi \uparrow$	$S_\alpha \uparrow$
✓		0.778	0.085	0.839	0.802	0.844	0.051	0.922	0.852	0.719	0.066	0.847	0.789
	✓	0.769	0.091	0.833	0.783	0.846	0.052	0.920	0.846	0.727	0.062	0.858	0.786
✓	✓	0.789	0.087	0.841	0.794	0.857	0.050	0.923	0.851	0.750	0.060	0.865	0.793

Table 17 Performance with different dilation sizes on ECSSD and DUT-OMRON datasets. The best results are highlighted in **bold**.

Dilation sizes	ECSSD				DUT-OMRON			
	$F_\beta \uparrow$	MAE $\downarrow$	$E_\phi \uparrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	MAE $\downarrow$	$E_\phi \uparrow$	$S_\alpha \uparrow$
(1,3,5,7)	0.872	0.060	0.900	0.857	0.698	0.068	0.842	0.771
(1,6,9,12)	0.857	0.069	0.897	0.870	0.658	0.079	0.810	0.780
(1,6,12,18)	0.862	0.061	0.895	0.855	0.685	0.075	0.829	0.767
Ours	0.878	0.056	0.908	0.864	0.708	0.066	0.844	0.778

Table 18 Performance with different dilation sizes on PASCAL-S, HKU-IS and DUTS-TE datasets. The best results are highlighted in **bold**.

Dilation sizes	PASCAL-S				HKU-IS				DUTS-TE			
	$F_\beta \uparrow$	MAE $\downarrow$	$E_\phi \uparrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	MAE $\downarrow$	$E_\phi \uparrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	MAE $\downarrow$	$E_\phi \uparrow$	$S_\alpha \uparrow$
(1,3,5,7)	0.779	0.090	0.834	0.786	0.857	0.052	0.920	0.845	0.748	0.061	0.860	0.788
(1,6,9,12)	0.763	0.095	0.826	0.805	0.832	0.062	0.915	0.859	0.694	0.073	0.826	0.799
(1,6,12,18)	0.767	0.094	0.828	0.785	0.842	0.054	0.915	0.844	0.723	0.068	0.847	0.781
Ours	0.789	0.087	0.841	0.794	0.857	0.050	0.923	0.851	0.750	0.060	0.865	0.793

PASCAL-S dataset) than the flipping operation. Third, when combined these transformation, the network can achieve the best results. For example, by introducing the single flipping transformation into the baseline, our method outperforms by 4.7% on average in terms of F-measure across the five datasets. Overall, these results strongly demonstrate that the consistent transformation mechanism can work well on our method.

Effect on different modules. To investigate the effectiveness of the proposed CAF module and ASPP module, we report F-measure, MAE, E-measure and S-measure values in Table 15 and Table 16 on five datasets. On the one hand, compared with w/o. CAF module, our model achieves the best performance enhancements in terms of all metrics, except the MAE score with the PASCAL-S dataset which shows decreased performance. Specifically, we find a significant drop in performance (DUT-OMRON: from 0.844 to 0.824, DUTS-TE: from 0.865 to 0.847) when using the E-measure. On the other hand, we can find that w/o. ASPP module provides the downtrend where the S-measure decreases from 0.851 to 0.846 on HKU-IS dataset. Furthermore,

some visual examples are shown in Fig. 14. We can observe that our CAF module and ASPP module can help to obtain complete structure information of the salient objects from the background.

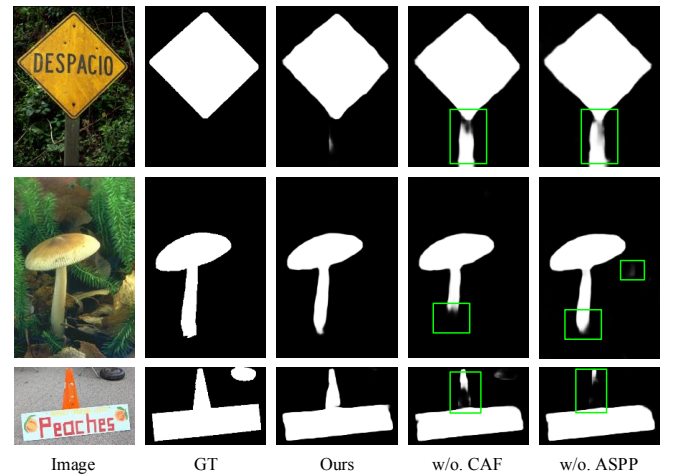
**Fig. 14** Visual analysis of different models.

Table 19 Ablation studies for various loss function on ECSSD and DUT-OMRON datasets. The best results are highlighted in **bold**.

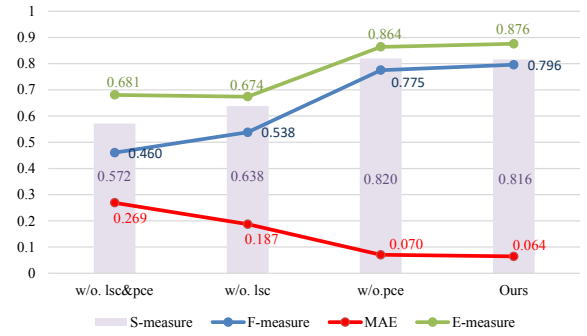
L_{pce}	L_{lsc}	L_{con}	L_{pse}	ECSSD				DUT-OMRON			
				$F_\beta \uparrow$	MAE $\downarrow$	$E_\phi \uparrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	MAE $\downarrow$	$E_\phi \uparrow$	$S_\alpha \uparrow$
✓		✓		0.551	0.268	0.709	0.612	0.374	0.266	0.657	0.542
✓		✓	✓	0.639	0.167	0.744	0.694	0.443	0.205	0.617	0.594
✓	✓	✓		0.865	0.064	0.903	0.866	0.673	0.074	0.822	0.777
✓	✓	✓	✓	0.878	0.056	0.908	0.864	0.708	0.066	0.844	0.778

Table 20 Ablation studies for various loss function on PASCAL-S, HKU-IS and DUTS-TE datasets. The best results are highlighted in **bold**.

L_{pce}	L_{lsc}	L_{con}	L_{pse}	PASCAL-S				HKU-IS				DUTS-TE			
				$F_\beta \uparrow$	MAE $\downarrow$	$E_\phi \uparrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	MAE $\downarrow$	$E_\phi \uparrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	MAE $\downarrow$	$E_\phi \uparrow$	$S_\alpha \uparrow$
✓		✓		0.473	0.284	0.653	0.571	0.503	0.258	0.721	0.591	0.398	0.267	0.666	0.545
✓		✓	✓	0.586	0.202	0.679	0.647	0.577	0.177	0.722	0.659	0.445	0.195	0.612	0.594
✓	✓	✓		0.775	0.090	0.833	0.803	0.845	0.056	0.920	0.856	0.715	0.066	0.841	0.797
✓	✓	✓	✓	0.789	0.087	0.841	0.794	0.857	0.050	0.923	0.851	0.750	0.060	0.865	0.793

Effect on dilation sizes in ASPP module. To validate the effectiveness of ASPP module, we replace the ASPP module with other dilation sizes, as shown in Table 17 and Table 18. First, it is clear that the dilation sizes (1,3,6,9) better than (1,3,5,7). Especially, the S-measure has been improved to different degrees, such as ECSSD dataset: 0.857→0.864; PASCAL-S dataset: 0.786→0.794, which means that our saliency maps have higher local and global similarity with the ground-truth. Second, when increasing the dilation sizes, the performance of saliency map is not continuous increase, setting dilation sizes as (1,6,9,12) or (1,6,12,18) show relatively low performance. For example, F-measure decreased 2.4% and 1.1% on ECSSD, respectively. These findings indicate that our design of ASPP module is useful to extract multi-scale context information for better saliency inference.

Effect on loss function. Here we analyze the validity of different key loss functions in Table 19 and Table 20. Accordingly, the first scheme (row 1) means that our network is a teacher-student model and the saliency loss only adopts PCE loss. The experimental results indicate that it has the lowest scores compared with other schemes. While the prediction results of the third scheme (row 3) are superior due to the inclusion of LSC loss, they are not as good as the fourth scheme (row 4) which includes communication between students. For example, the F-measure improves 1.5% and 4.9% on ECSSD and PASCAL-S datasets, respectively. Compared to our framework, the second scheme (row 2) changed the saliency loss L_{lsc} to test the effectiveness. It can be demonstrated that our approach produces more precise saliency maps. For fair comparison, we also present the average index value of our approach with others in Fig. 15. Benefitting from the good performance of S-measure on five datasets when

**Fig. 15** Ablation study of different loss function.

w/o. L_{pce} , thus it has the optimal S-measure. Generally speaking, although the S-measure of our proposed model is not the best, it still achieves relatively best quantitative performance compared with other methods.

Effect on seminar learning. To demonstrate the effectiveness of our proposed seminar learning, as shown in Table 21 and Table 22, we conduct the experiments on SHNet. "SHNet-SL" denotes the SHNet with seminar learning. At the same time, the specific strategy of seminar learning is consistent with the method of this paper. Specifically, it is observed that introducing seminar learning to SHNet can improve the performance. For example, compared with SHNet, the F-measure of SHNet-SL improves 1.8%, 1.4%, 1.9%, 1.0% and 2.0%, the MAE drops 13.5%, 5.3%, 17.0%, 5.5% and 4.3% on five datasets, respectively. These results indicate that SHNet-SL is more closer to ground-truth than SHNet. Overall, the seminar learning is benefit for SHNet.

Table 21 Ablation studies of seminar learning on ECSSD and DUT-OMRON datasets. The best results are highlighted in **bold**.

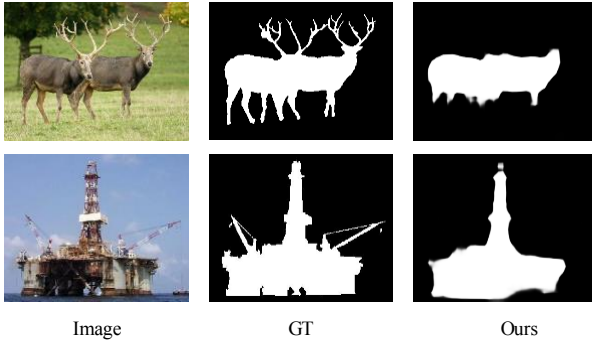
Method	ECSSD				DUT-OMRON			
	$F_\beta \uparrow$	MAE $\downarrow$	$E_\phi \uparrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	MAE $\downarrow$	$E_\phi \uparrow$	$S_\alpha \uparrow$
SHNet [43]	0.758	0.126	0.817	0.750	0.641	0.094	0.763	0.704
SHNet-SL	0.772	0.109	0.824	0.761	0.650	0.089	0.772	0.710

Table 22 Ablation studies of seminar learning on PASCAL-S, HKU-IS and DUTS-TE datasets. The best results are highlighted in **bold**.

Method	PASCAL-S				HKU-IS				DUTS-TE			
	$F_\beta \uparrow$	MAE $\downarrow$	$E_\phi \uparrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	MAE $\downarrow$	$E_\phi \uparrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	MAE $\downarrow$	$E_\phi \uparrow$	$S_\alpha \uparrow$
SHNet [43]	0.699	0.147	0.732	0.697	0.777	0.073	0.883	0.806	0.637	0.094	0.797	0.733
SHNet-SL	0.712	0.122	0.740	0.706	0.785	0.069	0.892	0.819	0.650	0.090	0.781	0.738

4.6 Failure cases

Fig. 16 present some failure cases on different scenes. The first line shows two deers on the grassland, our method is unable to distinguish the location of antlers. Similarly, the second row shows a machine in the image, it can be easily to observe that our method is insensitive to detail information processing. Overall, salient objects with more details may lead to incomplete segmentation.

**Fig. 16** Failure cases on different scenes.

4.7 Limitation

Although the results of the proposed method have reached the current optimal performance, there are still the following points worthy of further exploration. (1) It is not difficult to observe that our model is not lightweight (376.4 MB). (2) The proposed method can better locate and segment the approximate position of the salient object, but the performance on details are not clear and accurate. In the future, we will continue to explore these issues in our future work.

5 Conclusion

In this article, we present a seminar learning framework with consistent transformation ensembling (SLF-CT) for weakly supervised SOD. We first build a cross attention guided convolutional neural network for saliency predictions. Then we develop the teacher-student model and the student-student model to leverage scribble annotation to boost the performance of saliency prediction. A consistency transformation strategy is employed for further enhancing the regularization in the teacher-student model. To achieve the knowledge transfer between the student-student model, we apply multiple pseudo labels to each pixel that is able to alleviate the shortage of scribble annotations as well. Furthermore, our proposed SLF-CT with the scribble annotations outperforms state-of-the-art weakly supervised methods and unsupervised methods, and is able to achieve 92.8% accuracy of fully supervised methods at least.

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

References

1. Achanta, R., Hemami, S., Estrada, F., Susstrunk, S.: Frequency-tuned salient region detection. In: 2009 IEEE conference on computer vision and pattern recognition, pp. 1597–1604. IEEE (2009)
2. Beaulieu-Laroche, L., Toloza, E.H., Van der Goes, M.S., Lafourcade, M., Barnagian, D., Williams, Z.M., Eskandar, E.N., Frosch, M.P., Cash, S.S., Harnett, M.T.: Enhanced dendritic compartmentalization in human cortical neurons. *Cell* **175**(3), 643–651 (2018)
3. 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)
4. Chen, C., Song, J., Peng, C., Wang, G., Fang, Y.: A novel video salient object detection method via semisupervised motion quality perception. *IEEE Transactions on Circuits and Systems for Video Technology* **32**(5), 2732–2745 (2021)
5. Chen, C., Wei, J., Peng, C., Qin, H.: Depth-quality-aware salient object detection. *IEEE Transactions on Image Processing* **30**, 2350–2363 (2021)
6. Chen, G., Choi, W., Yu, X., Han, T., Chandraker, M.: Learning efficient object detection models with knowledge distillation. *Advances in neural information processing systems* **30** (2017)
7. Chen, H., Wang, J., Chen, H.C., Zhen, X., Zheng, F., Ji, R., Shao, L.: Seminar learning for click-level weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6920–6929 (2021)
8. 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)
9. Chen, T., Hu, X., Xiao, J., Zhang, G., Wang, S.: Binet: Bidirectional interactive network for salient object detection. *Neurocomputing* **465**, 490–502 (2021)
10. 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)
11. 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)
12. Ding, W., Abdel-Basset, M., Hawash, H.: Rctc: A reliable and consistent temporal-ensembling framework for semi-supervised segmentation of covid-19 lesions. *Information sciences* **578**, 559–573 (2021)
13. 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)
14. 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)
15. 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)
16. Fang, X., Zhu, J., Zhang, R., Shao, X., Wang, H.: Ibbnet: Interactive branch network for salient object detection. *Neurocomputing* **465**, 574–583 (2021)
17. Goferman, S., Zelnik-Manor, L., Tal, A.: Context-aware saliency detection. *IEEE transactions on pattern analysis and machine intelligence* **34**(10), 1915–1926 (2011)
18. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778 (2016)
19. 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)
20. 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)
21. 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)
22. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7132–7141 (2018)
23. 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)
24. Li, G., Xie, Y., Lin, L.: Weakly supervised salient object detection using image labels. In: Thirty-second AAAI conference on artificial intelligence (2018)
25. 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)
26. Li, G., Yu, Y.: Visual saliency detection based on multiscale deep cnn features. *IEEE transactions on image processing* **25**(11), 5012–5024 (2016)
27. Li, H., Li, G., Yang, B., Chen, G., Lin, L., Yu, Y.: Depth-wise nonlocal module for fast salient object detection using a single thread. *IEEE Transactions on Cybernetics* (2020)
28. 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)
29. Li, X., Yu, L., Chen, H., Fu, C.W., Xing, L., Heng, P.A.: Transformation-consistent self-ensembling model for semisupervised medical image segmentation. *IEEE Transactions on Neural Networks and Learning Systems* **32**(2), 523–534 (2020)
30. Lin, D., Dai, J., Jia, J., He, K., Sun, J.: Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3159–3167 (2016)
31. Liu, J.J., Hou, Q., Cheng, M.M.: Dynamic feature integration for simultaneous detection of salient object, edge, and skeleton. *IEEE Transactions on Image Processing* **29**, 8652–8667 (2020)
32. 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)

33. Liu, J.J., Hou, Q., Liu, Z.A., Cheng, M.M.: Poolnet+: Exploring the potential of pooling for salient object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022)
34. Liu, J.J., Liu, Z.A., Peng, P., Cheng, M.M.: Rethinking the u-shape structure for salient object detection. *IEEE Transactions on Image Processing* **30**, 9030–9042 (2021)
35. Liu, N., Zhang, N., Wan, K., Han, J., Shao, L.: Visual saliency transformer. *arXiv preprint arXiv:2104.12099* (2021)
36. Liu, Q., Yu, L., Luo, L., Dou, Q., Heng, P.A.: Semi-supervised medical image classification with relation-driven self-ensembling model. *IEEE transactions on medical imaging* **39**(11), 3429–3440 (2020)
37. Liu, X., Yuan, Q., Gao, Y., He, K., Wang, S., Tang, X., Tang, J., Shen, D.: Weakly supervised segmentation of covid19 infection with scribble annotation on ct images. *Pattern recognition* **122**, 108341 (2022)
38. Liu, Y., Cheng, M.M., Zhang, X.Y., Nie, G.Y., Wang, M.: DNA: Deeply supervised nonlinear aggregation for salient object detection. *IEEE Transactions on Cybernetics* **52**(7), 6131–6142 (2022)
39. 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)
40. 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)
41. 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)
42. Ma, G., Chen, C., Li, S., Peng, C., Hao, A., Qin, H.: Salient object detection via multiple instance joint re-learning. *IEEE Transactions on Multimedia* **22**(2), 324–336 (2019)
43. Ma, L., Luo, X., Hong, H., Zhang, Y., Wang, L., Wu, J.: Scribble-attention hierarchical network for weakly supervised salient object detection in optical remote sensing images. *Applied Intelligence* pp. 1–19 (2022)
44. 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)
45. 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)
46. 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)
47. 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)
48. Piao, Y., Wu, W., Zhang, M., Jiang, Y., Lu, H.: Noise-sensitive adversarial learning for weakly supervised salient object detection. *IEEE Transactions on Multimedia* (2022)
49. Samuli, L., Timo, A.: Temporal ensembling for semi-supervised learning. In: *International Conference on Learning Representations (ICLR)*, vol. 4, p. 6 (2017)
50. Shi, X., Su, H., Xing, F., Liang, Y., Qu, G., Yang, L.: Graph temporal ensembling based semi-supervised convolutional neural network with noisy labels for histopathology image analysis. *Medical image analysis* **60**, 101624 (2020)
51. Siagian, C., Itti, L.: Rapid biologically-inspired scene classification using features shared with visual attention. *IEEE transactions on pattern analysis and machine intelligence* **29**(2), 300–312 (2007)
52. 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)
53. Tang, M., Perazzi, F., Djelouah, A., Ben Ayed, I., Schroers, C., Boykov, Y.: On regularized losses for weakly-supervised cnn segmentation. In: *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 507–522 (2018)
54. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems* **30** (2017)
55. 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)
56. 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)
57. Wang, W., Zhao, S., Shen, J., Hoi, S.C., Borji, A.: Salient object detection with pyramid attention and salient edges. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1448–1457 (2019)
58. Wang, X., Al-Huda, Z., Peng, B.: Weakly-supervised salient object detection through object segmentation guided by scribble annotations. In: *2021 16th International Conference on Intelligent Systems and Knowledge Engineering (ISKE)*, pp. 304–312. *IEEE* (2021)
59. Wang, Z., Zhang, Y., Liu, Y., Liu, S., Coleman, S., Kerr, D.: Mfc-net: Multi-feature fusion cross neural network for salient object detection. *Image and Vision Computing* p. 104243 (2021)
60. Wei, J., Wang, S., Huang, Q.: F³net: Fusion, feedback and focus for salient object detection. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 12321–12328 (2020)
61. 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)
62. Wu, Z., Li, S., Chen, C., Qin, H., Hao, A.: Salient object detection via dynamic scale routing. *IEEE Transactions on Image Processing* **31**, 6649–6663 (2022)
63. 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)
64. Wu, Z., Wang, L., Wang, W., Shi, T., Chen, C., Hao, A., Li, S.: Synthetic data supervised salient object detection. In: *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 5557–5565 (2022)

65. Yan, K., Wang, X., Kim, J., Feng, D.: A new aggregation of dnn sparse and dense labeling for saliency detection. *IEEE Transactions on Cybernetics* **51**(12), 5907–5920 (2020)
66. Yan, P., Li, G., Xie, Y., Li, Z., Wang, C., Chen, T., Lin, L.: Semi-supervised video salient object detection using pseudo-labels. In: *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7284–7293 (2019)
67. 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)
68. 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)
69. Yang, W., Zhang, T., Yu, X., Qi, T., Zhang, Y., Wu, F.: Uncertainty guided collaborative training for weakly supervised temporal action detection. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 53–63 (2021)
70. Yu, S., Zhang, B., Xiao, J., Lim, E.G.: Structure-consistent weakly supervised salient object detection with local saliency coherence. In: *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*. AAAI Palo Alto, CA, USA (2021)
71. 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)
72. 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)
73. 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)
74. 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)
75. 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)
76. 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)
77. Zhang, P., Wang, D., Lu, H., Wang, H., Ruan, X.: A-mulet: Aggregating multi-level convolutional features for salient object detection. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 202–211 (2017)
78. 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)
79. 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)
80. 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)
81. 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)
82. Zhuge, M., Fan, D.P., Liu, N., Zhang, D., Xu, D., Shao, L.: Salient object detection via integrity learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022)