Personal Fixations-Based Object Segmentation With Object Localization and Boundary Preservation

Gongyang Li, *Member, IEEE*, Zhi Liu, *Senior Member, IEEE*, Ran Shi, Zheng Hu, Weijie Wei, Yong Wu, Mengke Huang, and Haibin Ling

***Abstract*— As a natural way for human-computer interaction, fixation provides a promising solution for interactive image segmentation. In this paper, we focus on Personal Fixations-based Object Segmentation (PFOS) to address issues in previous studies, such as the lack of appropriate dataset and the ambiguity in fixations-based interaction. In particular, we first construct a new PFOS dataset by carefully collecting pixel-level binary annotation data over an existing fixation prediction dataset, such dataset is expected to greatly facilitate the study along the line. Then, considering characteristics of personal fixations, we propose a novel network based on Object Localization and Boundary Preservation (OLBP) to segment the gazed objects. Specifically, the OLBP network utilizes an Object Localization Module (OLM) to analyze personal fixations and locates the gazed objects based on the interpretation. Then, a Boundary Preservation Module (BPM) is designed to introduce additional boundary information to guard the completeness of the gazed objects. Moreover, OLBP is organized in the mixed bottom-up and top-down manner with multiple types of deep supervision. Extensive experiments on the constructed PFOS dataset show the superiority of the proposed OLBP network over 17 state-of-the-art methods, and demonstrate the effectiveness of the proposed OLM and BPM components. The constructed PFOS dataset and the proposed OLBP network are available at https://github.com/MathLee/OLBPNet4PFOS.**

***Index Terms*— Personal fixations, interactive image segmentation, object localization, boundary preservation.**

I. Introduction

FIXATION is a flexible interaction mechanism of the human visual system. Compared with scribble, click and bounding box, fixation provides the most convenient interaction for patients with hand disability, amyotrophic lateral sclerosis (ALS) and polio. This kind of eye control interaction, *i.e*. fixation, can greatly improve the interaction efficiency of these patients. In addition, fixation is closely related to personal information such as age [1], [2] and gender [3], [4]. This means that different individuals may have different perceptions and preferences of a scene [5], [6]. Thus motivated, in this paper, we pay close attention to personal fixations-based object segmentation, which is a more natural manner for interactive image segmentation.

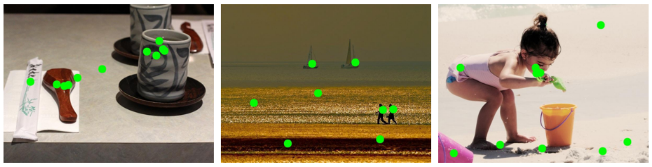


Fig. 1. Examples of image with ambiguous fixations. Green dots in each image indicate fixations. Some fixations fall in the background.

The typical manners of interaction, such as scribbles [7]–[11], clicks [12]–[17] and bounding boxes [18]–[21] for interactive image segmentation, are explicit behaviors without interference. By contrast, fixations are implicit [22]–[25], and their convenience comes with interaction ambiguity. Concretely, the positive and negative labels of scribbles and clicks are deterministic. However, fixations are unlabeled when collected. They do not distinguish between positive labels and negative labels (*i.e*. some fixations may fall in the background as shown in Fig. 1), resulting in a few noise in the fixations. Such ambiguous interaction makes the fixations-based object segmentation task difficult. Recently, with the rise of convolutional neural networks (CNNs), the clicks-based interactive image segmentation has been greatly developed. Even though fixation points and clicking points are similar to some extent, clicks-based methods [12]–[14], [16], [17] cannot be directly applied to fixations-based object segmentation.

The above observations suggest that there are two main reasons that limit the development of fixations-based object segmentation. First, there is not a suitable dataset for the fixations-based object segmentation task, let alone dataset based on the personal fixations. Second, as aforementioned, the ambiguous representation of fixations makes this type of interaction difficult to handle by other methods which are based on clicks and scribbles.

To address the first crucial issue, we construct a *Personal Fixations-based Object Segmentation* (PFOS) dataset, which is extended from the fixation prediction dataset OSIE [26].

Manuscript received April 21, 2020; revised October 19, 2020; accepted December 8, 2020. Date of publication December 18, 2020; date of current version December 31, 2020. This work was supported by the National Natural Science Foundation of China under Grant 61771301 and Grant 61801219. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Shuicheng Yan. *(Corresponding author: Zhi Liu.)*

Gongyang Li, Zhi Liu, Zheng Hu, Weijie Wei, Yong Wu, and Mengke Huang are with Shanghai Institute for Advanced Communication and Data Science, Shanghai University, Shanghai 200444, China, and also with the School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China (e-mail: ligongyang@shu.edu.cn; liuzhisjtu@163.com; huzhen1995@shu.edu.cn; codename1995@shu.edu.cn; yong_wu@shu.edu.cn; huangmengke@shu.edu.cn).

Ran Shi is with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: rshi@njust.edu.cn).

Haibin Ling is with the Department of Computer Science, Stony Brook University, Stony Brook, NY 11794 USA (e-mail: hling@cs.stonybrook.edu).

Digital Object Identifier 10.1109/TIP.2020.3044440

The PFOS dataset contains 700 images, and each image has 15 personal fixation maps collected from 15 subjects with corresponding pixel-level annotations of objects. To overcome the ambiguity of fixations, we propose an effective network based on *Object Localization and Boundary Preservation* (OLBP). The key idea of OLBP is to locate the gazed objects based on the analysis of fixations, and then the boundary information is introduced to guard the completeness of the gazed objects and to filter the background.

In particular, the overall structure of OLBP network is a mixture of bottom-up and top-down architectures. To narrow the gap between fixations and objects, we propose the *Object Localization Module* (OLM) to analyze personal fixations in detail and grasp location information of the gazed objects of different individuals. Based on the interpretation of location information, OLM modulates CNN features of image in a bottom-up way. Moreover, considering that the object location information may involve confusing noise, we propose a *Boundary Preservation Module* (BPM) to exploit boundary information to enforce object completeness and filter the background of erroneous localization. BPM is integrated into the top-down prediction. Both OLMs and BPMs employ deep supervision to further improve the capabilities of feature representation. In this way, the scheme of object localization and boundary preservation is successfully applied to the bottom-up and top-down structure, and the proposed OLBP network greatly promotes the performance of the personal fixations-based object segmentation task. Experimental results on the challenging PFOS dataset demonstrate that OLBP outperforms 17 state-of-the-art methods under various evaluation metrics.

The contributions of this work are summarized as follows:

- We construct a new dataset for *Personal Fixations-based Object Segmentation* (PFOS), which focuses on the natural interaction (*i.e.* fixation). This dataset contains free-view personal fixations without any constraints, expanding its applicability. We believe that the PFOS dataset will boost the research of fixations-based human-computer interaction.
- We propose a novel *Object Localization and Boundary Preservation* (OLBP) network to segment the gazed objects based on personal fixations. The OLBP network, equipped with the *Object Localization Module* and the *Boundary Preservation Module*, effectively overcomes the difficulties from ambiguous fixations.
- We conduct extensive experiments to evaluate our OLBP network and other state-of-the-art methods on the PFOS dataset. Comprehensive results demonstrate the superiority of our OLBP network, and also reveal the difficulties and challenges of the constructed PFOS dataset.

The rest of the paper is organized as follows: Sec. II reviews related previous works. Then, we formulate the PFOS task in Sec. III. After that, in Sec. IV, we construct the PFOS dataset. Sec. V presents the proposed OLBP network in detail. In Sec. VI, we evaluate the performance of the proposed OLBP network and other methods on the constructed PFOS dataset. Finally, the conclusion is drawn in Sec. VII.

II. Related Work

In this section, we first give an overview of previous works of interactive image segmentation in Sec. II-A. Then, we introduce related works of fixations-based object segmentation in Sec. II-B. Finally, we review some related works on boundary-aware segmentation in Sec. II-C.

A. Interactive Image Segmentation

1) Scribbles-Based Interactive Image Segmentation: Scribble is a traditional manner of interaction. Most of scribbles-based methods are built on graph structures. Graph-Cut [7] is one of the most representative methods. It uses the max-flow/min-cut theorem to minimize energy function with hard constraints (*i.e.* labeled scribbles) and soft constraints. Grady *et al.* [8] adopted the random walk algorithm to assign a label to each unlabeled pixel based on the predefined seed pixels in discrete space. In [9], Bai *et al.* proposed a weighted geodesic distance based framework, which is fast for image and video segmentation and matting. Nguyen *et al.* [10] proposed a convex active contour model to segment objects, and their results were with smooth and accurate boundary contour. Spina *et al.* [11] presented a live markers methodology to reduce the user intervention for effective segmentation of target objects. Following the seed propagation strategy, Jian *et al.* [27] employed the adaptive constraint propagation to adaptively propagate the scribbles information into the whole image. Recently, Wang *et al.* [28] changed their view on interactive image segmentation and formulated it as a probabilistic estimation problem, proposing a pairwise likelihood learning based framework. These methods are friendly to clearly defined scribbles, but they cannot solve the ambiguity of fixations and their inference speed is usually slow.

2) Clicks-Based Interactive Image Segmentation: Click is a classical manner of interaction. It has been deeply studied in the deep learning era. The positive and negative clicks are transformed into two separate Euclidean distance maps for network input. Xu *et al.* [12] directly sent RGB image and two distance maps into a fully convolutional network. Liew *et al.* [13] proposed a two-branch fusion network with global prediction and local regional refinement. In addition to the RGB image and distance maps, Li *et al.* [14] included clicks in their network input and proposed an end-to-end segmentation-selection network. In [16], Jang *et al.* introduced the backpropagating refinement scheme to correct mislabeled locations in the initial segmentation map. Different from the direct concatenation of RGB image and interaction maps of the above methods, Hu *et al.* [17] separately input RGB image and interaction maps into two networks, and designed a fusion network for feature interactions. CNNs have greatly improved the performance of clicks-based interactive image segmentation, but when these methods are applied to fixations-based object segmentation, some background regions will be mistakenly segmented. To address the problem of erroneous localization, we explore the boundary information in our BPM to filter redundant background regions and guard the gazed object.

3) Bounding Boxes-Based Interactive Image Segmentation: In a bounding box, the target object and background coexist, which is different from scribble and click. Rother *et al.* [18] extended the graph-cut approach, and segmented object with a rectangle, namely GrabCut. To overcome the looseness of the bounding box, Lempitsky *et al.* [19] incorporated the tightness prior into the global energy minimization function as hard constraints to further completed target object. Shi *et al.* [21] proposed a coarse-to-fine method with region-level and pixel-level segmentation. Similar to [12], Xu *et al.* [20] transformed the bounding box to a distance map and concatenated it with the RGB image to input into an encoder-decoder network. Although bounding box and fixation are similar (*i.e.* target object and background coexist in both interactions), the bounding box-based methods are difficult to transfer to fixations-based object segmentation.

B. Fixations-Based Object Segmentation

Fixation plays an integral role in the human visual system and it is convenient for interaction. In an early study, Sadeghi *et al.* [29] constructed an eyegaze-based interactive segmentation system which adopts random walker to segment objects. Meanwhile, Mishra *et al.* [22] gave the definition of fixations-based object segmentation, that is, segmenting regions containing fixation points. They transformed the image to polar coordinate system, and found the optimal contour to fit the target object. Based on the interpretation of visual receptive field, Kootstra *et al.* [30] used symmetry to select fixations closer to the center of the object to obtain more complete segmentation. Differently, Li *et al.* [23] focused on selecting the most salient objects, and they ranked object proposals based on fixations. Similar to [23], Shi *et al.* [24] analyzed the fixation distribution and proposed three metrics to evaluate the score of each candidate region. In [31], Tian *et al.* first determined the uninterested regions, and then used superpixel-based random walk model to segment the gazed objects. Khosravan *et al.* [32] integrated fixations into the medical image segmentation and proposed a Gaze2Segment system. Li *et al.* [25] constructed a dataset where all fixations fall in objects (*i.e.* constrained fixations), and proposed a CNN-based model to simulate the human visual system to segment objects based on fixations.

These studies have promoted the development of fixations-based object segmentation. However, all the fixations in [22], [25], [30], [31] fall in objects, which are hardly guaranteed in practice. These methods [22], [25], [30], [31] will get stuck in the ambiguity of unconstrained fixations, especially of personal fixations. For [23], [24], they are based on region proposal and cannot obtain accurate results. In summary, the above methods cannot solve the problem of ambiguous fixations, as shown in Fig. 1. In this paper, we take advantage of CNNs, and propose a bottom-up and top-down network to locate objects and preserve objects' boundaries. Moreover, we construct a dataset to promote this special direction of interactive image segmentation, *i.e.* personal fixations-based object segmentation.

C. Boundary-Aware Segmentation

The boundary/edge-aware segmentation idea is widely-used in salient object detection [33]–[36] and semantic segmentation [37]. In [33], Wang *et al.* modeled the boundary information as an edge-preserving constraint, and included it as an additional supervision in loss function. In [34], Wang *et al.* proposed a two-branch network, including boundary and mask sub-networks, for jointly predicting masks of salient objects and detecting object boundaries. In [35], Wu *et al.* explored the logical interrelations between binary segmentation and edge maps in a multi-task network, and proposed a cross refinement unit in which the segmentation features and edge features are fused in a cross-task manner. In [36], Zhao *et al.* focused on the complementarity between salient edge information and salient object information. They integrated the local edge information of shallow layers and global location information of deep layers to obtain the salient edge features, and then the edge features were fed to the one-to-one guidance module to fuse the complementary region and edge information. In [37], Ding *et al.* first introduced the boundary information as an additional semantic class to enable the network to be aware of the boundary layout, and then proposed a boundary-aware feature propagation network to control the feature propagation based on the learned boundary information.

In our method, we use the boundary information in two aspects: the multi-task structure (*i.e.* segmentation and boundary predictions) and the *Boundary Preservation Module*. Different from [34], [35], we integrate the learned boundary map into the prediction network in BPMs to preserve the completeness of the gazed objects, rather than fuse the segmentation features and boundary features. Compared with [36], our segmentation prediction is accompanied by the boundary prediction in a uniform prediction network, and the boundary supervision is employed at multiple scales. Different from [37], which uses the boundary map to control the region of feature propagation, our method uses the boundary map to filter the background of erroneous localization in features. In short, our use of boundary information is diverse and in-depth, which is suitable for the personal fixations-based object segmentation task.

III. Personal Fixations-Based Object Segmentation

A. Problem Statement

Given an image **I** and a fixation map **FM** of a person, personal fixations-based object segmentation aims to segment the gazed objects of this person according to his/her personal **FM**, producing a binary segmentation map. In general, different individuals generate different fixation maps when observing the same image, which means that individuals may be interested in different objects. In other words, segmentation results of different individuals on the same image vary with the observer. So, the special characteristic of this task is that an image has multiple binary segmentation maps due to multiple fixation maps. Although the ambiguity of fixations makes this task difficulty, the personal fixation map is the only information that can determine the gazed objects.

B. Applications

This task has several meaningful applications. First, such a convenient manner of interaction is conducive to the development of special eye-control devices for patients with hand disability, ALS and polio, facilitating their lives and improving their quality of life. Second, fixation is advantageous to diagnose certain mental illnesses, such as autism spectrum disorder (ASD) [38], [39] and schizophrenia spectrum disorders (SSD) [40], [41]. This task understands personal fixations at the object level, which is helpful to improve the accuracy of disease diagnosis. For example, patients with ASD prefer to pay attention to background rather than foreground, so the proportion of foreground in their segmentation results will be less than that of healthy people.

IV. Dataset Construction and Transformation

Currently, there are many prevalently used datasets for fixation prediction, such as MIT1003 [42], OSIE [26] and SALICON [43], and for interactive image segmentation, such as GrabCut [18], Berkeley [44] and PASCAL VOC [45]. However, there is no dataset for the personal fixations-based object segmentation task. Considering that it is time-consuming for dataset annotations, we propose a convenient way to collect suitable data from existing datasets for this task.

Obviously, the PFOS dataset must contain fixation data and pixel-level annotations for objects. Among the existing datasets, some datasets, such as DUTS-OMRON [46], PASCAL-S [23] and OSIE [26], are potential candidates. The pixel-level annotations of DUTS-OMRON and PASCAL-S are for salient object detection [47]–[49], that is, these annotations only focus on the most visually attractive objects but ignore other objects, which could be fixated by different individuals, in a scene. Therefore, they are not perfect for constructing a PFOS dataset. Fortunately, the pixel-level annotations of OSIE have semantic attributes. This means that we can select objects, which the user is interested in, based on personal fixations. In other words, we can create the pixel-level binary ground truths (GTs) for personal fixations-based object segmentation. So, we transform the fixation prediction dataset OSIE to our PFOS dataset.

For each image in the OSIE dataset, it has corresponding fixation maps and semantic GTs of different subjects. The detailed steps for dataset transformation are as follows:

1) *Semantic labels collection.* We get the position of each fixation point from the fixation map, and we collect the semantic label of each position in the corresponding semantic GT.

2) *Semantic labels distillation.* As mentioned in Sec. I and shown in Fig. 1, some fixation points fall in the background or the same object. For semantic labels collected from Step 1, we discard the semantic label "0" which indicates background. Then, if there are several same semantic labels, we keep only one.

3) *Binary GT creation.* Based on the distilled semantic labels from Step 2, we can determine the gazed objects and create the binary GT. We reserve the regions with the distilled semantic labels in the semantic GT, and set them as foreground. We set the regions with the other unrelated semantic labels as background.

TABLE I

Categories of Fixation Map (FM) in the PFOS Dataset. Constrained FM Means That All Fixations Fall in the Objects/Foreground. Unconstrained FM Represents That Some Fixations Fall in the Background

PFOS dataset	Constrained FM	Unconstrained FM
10,500	3,683 (35.1%)	6,817 (64.9%)

In this convenient way, we efficiently create the binary GTs and successfully construct the PFOS dataset. The PFOS dataset retains all 700 images and 10,500 free-view personal fixation maps from the OSIE dataset. In the PFOS dataset, the image resolution is 800×600. Each image has 15 personal fixation maps from 15 subjects and the transformed binary GTs. In the constructed PFOS dataset, there are two categories of fixation maps. The first category is that all fixations fall in the objects/foreground, *i.e.* the constrained fixation map in [25]. The second category is that some fixations fall in the background, namely the unconstrained fixation map. We present the details of them in Tab. I. In our PFOS dataset, the unconstrained fixation maps account for 64.9% and the constrained fixation maps hold 35.1%. The large proportion of unconstrained fixation maps increase the ambiguity of our PFOS dataset and make this dataset challenging.

V. Methodology

In this section, we first conduct data preprocessing which transforms the fixation points into fixation density maps in Sec. V-A. Then, we present the overview and motivation of the proposed *Object Localization and Boundary Preservation* (OLBP) network in Sec. V-B. Next, we give the detailed formulas of the *Object Localization Module* (OLM) and the *Boundary Preservation Module* (BPM) in Sec. V-C and Sec. V-D, respectively. Finally, we clarify the implementation details of OLBP network in Sec. V-E.

A. Data Preprocessing

The fixation points in each fixation map are sparse. With only a few pixels per fixation map, there is too little valuable information to supply. The similar problem arises in the clicks-based interactive image segmentation. Xu *et al.* [12] transformed the clicks into Euclidean distance maps. Inspired by this, we employ the Gaussian blur to transform the sparse fixation map (*i.e.* FM) into the *fixation density map* (*i.e.* FDM):

$$\mathbf{FDM} = \mathrm{nor}_{\mathrm{min-max}}(\mathbf{FM} \circledast G_{\sigma}(x, y; \sigma)), \tag{1}$$

where $\mathrm{nor}_{\mathrm{min-max}}(\cdot)$ is the min-max normalization, $\circledast$ denotes convolution operator, and $G_{\sigma}(\cdot)$ is a Gaussian filter with parameter σ which is the standard deviation. σ is set corresponding to 1° visual angle in the OSIE dataset [26]. It is 24 pixels of an 800×600 image by default.

The effect of Gaussian blur is similar to the receptive field of eye, that is, the center of fixation is with a high resolution and

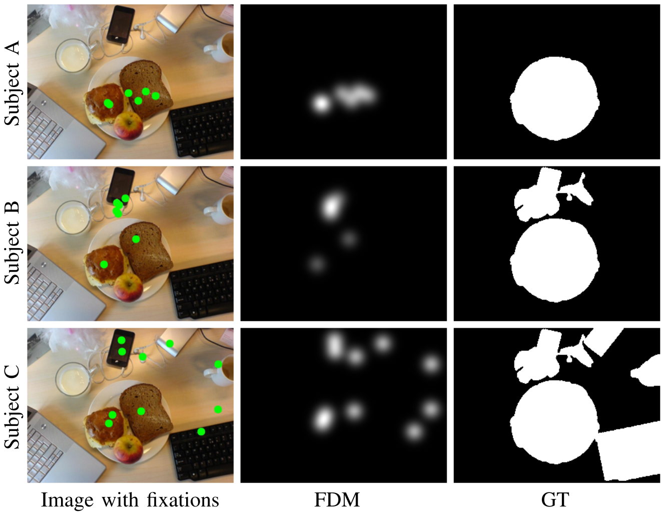


Fig. 2. Examples of the PFOS dataset. Green dots in each image indicate fixations, FDM is fixation density map, and GT represents ground truth.

the surrounding of fixation is with a low resolution. Thus, after performing Gaussian blur and linear transformation on FM, the dense FDM contains more prior information of objects. In this paper, we adopt the dense FDM rather than the raw FM. We present an image with the personal fixations of three subjects of the PFOS dataset in Fig. 2. The fixation maps of Subject A and Subject B are constrained fixation maps, while the fixation map of Subject C is an unconstrained fixation map.

B. Network Overview and Motivation

The proposed OLBP network has three critical components: the feature extractor, the object locator and the prediction network with boundary preservation. The overall architecture of OLBP network is illustrated in Fig. 3.

1) Feature Extractor: In the OLBP network, we adopt the modified VGG-16 [50], from which the last three fully connected layers have been removed, as the feature extractor. We denote its input image as $\mathbf{I} \in \mathbb{R}^{H\times W\times C}$, and initialize its parameters by the image classification model [50]. The feature extractor has five convolutional blocks, as shown in Fig. 3. We operate on the feature map of the last convolutional layer in each block, *i.e. conv1-2*, *conv2-2*, *conv3-3*, *conv4-3* and *conv5-3*, which are denoted as $\{\mathbf{F}_r^{(i)}\colon \mathbf{F}_r^{(i)} \in \mathbb{R}^{h_i\times w_i\times c_i}, i=1,2,\ldots,5\}$. Notably, the feature resolution at the *i-th* block, *i.e.* $[h_i, w_i]$, is $[\frac{H}{2^{i-1}}, \frac{W}{2^{i-1}}]$ and $c_{i\in\{1,2,3,4,5\}} = \{64, 128, 256, 512, 512\}$. In reality, the input resolution $[H, W, C]$ of $\mathbf{I}$ is set to $288 \times 288 \times 3$.

2) Object Localization Module: Although FDM is a probability map, it is a critical interaction that reflects the intention of the user. It is important to effectively explore the object location information of FDM. However, when we construct a CNN-based model for the personal fixations-based object segmentation task, it is natural to directly concatenate FDM and the input image for the network input. Since there are three channels for image and only one channel for FDM, the direct concatenation operation may drown out the critical interaction information of FDM. Based on the above analysis, we propose the *Object Localization Module* to process FDM.

The parallel convolution structure is effective to explore meaningful information in CNN features [51], especially with the dilated convolution [52]. Thus, in OLM, we employ several parallel dilated convolutions with different dilation rates to profoundly analyze the personal FDM to obtain object location information, which are a group of response maps. These response maps belong to $[0, 1]^{h_i\times w_i\times c_i}$, which shows they have the same number of channels as the features of image at the *i-th* block. They are applied to re-weight features of image to highlight the gazed objects at channel-wise and spatial-wise. To enhance the location presentation of the response maps, we apply deep supervision [53] in OLM. As presented in Fig. 3, the OLM is performed in a bottom-up manner, and it is assembled after each block of feature extractor for strong object localization. The detailed description of OLM is presented in Sec. V-C. We show the ablation study of OLM in Sec. VI-C, including a variant of direct concatenation of image and FDM.

3) Boundary Preservation Module and Prediction Network: Since some fixations fall in the background, there may be some noise on the re-weighted feature of OLM. The ambiguity over the fixations causes great disturbance to the segmentation result. Fortunately, there is *a priori knowledge* that the background usually does not have a regular boundary. Thus, we introduce the boundary information into the prediction network, and propose the *Boundary Preservation Module* to filter the background of erroneous localization and preserve the completeness of the gazed objects. BPM is a momentous component to purify the segmentation result. We also attach the pixel-level segmentation supervision and boundary supervision to BPM. As shown in Fig. 3, BPMs are equipped between convolutional blocks in the prediction network from top to down. To make full use of the boundary information, we also construct a multi-task structure in the prediction network. We elaborate the formulation and ablation study of BPM in Sec. V-D and Sec. VI-C, respectively.

C. Object Localization Module

As the **OLM-5** shown in Fig. 3, there are three main parts in the *Object Localization Module*: location analysis unit, feature re-weighting (*i.e.* Re-wei) and segmentation supervision (*i.e.* Seg sup). Its objective is to extract object location information of personal FDM and to highlight objects in feature of image $\mathbf{F}_r^{(i)}$. OLM is the most indispensable part of the whole OLBP network.

Concretely, in OLM-*i*, the $\mathbf{FDM} \in \mathbb{R}^{H\times W\times 1}$ is first downsampled to fit the resolution of $\mathbf{F}_r^{(i)}$ and to generate $\mathbf{F}_{fdm}^{(i)} \in \mathbb{R}^{h_i\times w_i\times 1}$ which is formulated as:

$$\mathbf{F}_{fdm}^{(i)} = \mathrm{MaxPool}(\mathbf{FDM}; W_{ks}^{(i)}), \tag{2}$$

where MaxPool$(\cdot)$ is the max pooling with parameters $W_{ks}^{(i)}$, which are $2^{i-1} \times 2^{i-1}$ kernel with 2^{i-1} stride.

Then, we design the location analysis unit, which contains four parallel dilated convolutions [52] with different dilation rates, to analyze $\mathbf{F}_{fdm}^{(i)}$, and obtain the multi-interpretation

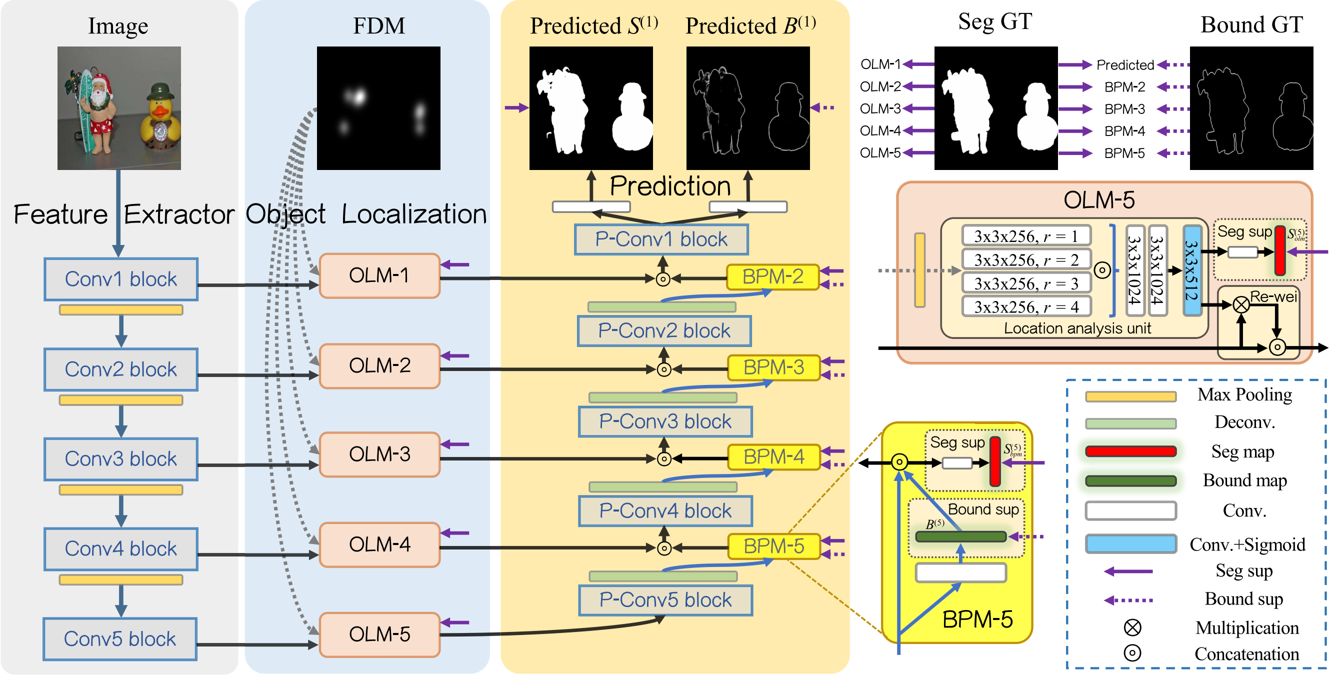


Fig. 3. The overall architecture of the proposed OLBP network. OLBP network is organized in the mixed bottom-up and top-down manner. We employ the modified VGG-16 to extract five blocks of features from an input image. Then in each OLM, FDM is analyzed by several dilated and normal convolutional layers to determine the location of objects in the corresponding block features. Based on the object localization in each feature block, the top-down prediction is established. During the prediction process, the boundary information is introduced into BPMs to guard the completeness of objects and to filter background of erroneous localization. We also construct a multi-task prediction structure, which contains object segmentation branch and boundary prediction branch, to exploit the complementarity between regions and boundaries.

feature $\mathbf{F}_{mi}^{(i)}$. The process in this unit can be formulated as:

$$\mathbf{F}_{mi}^{(i)} = \text{concat}\big(C_d(\mathbf{F}_{fdm}^{(i)}; W_d^{(i_1)}), C_d(\mathbf{F}_{fdm}^{(i)}; W_d^{(i_2)}), C_d(\mathbf{F}_{fdm}^{(i)}; W_d^{(i_3)}), C_d(\mathbf{F}_{fdm}^{(i)}; W_d^{(i_4)})\big), \tag{3}$$

where concat$(\cdot)$ is the cross-channel concatenation, and $C_d(\cdot; W_d^{(i_n)})$ is the dilated convolution with parameters $W_d^{(i_n)}$ for $n \in \{1, 2, 3, 4\}$. Notably, $W_d^{(i_n)}$ are comprised of kernel size, channel number and dilation rate. Considering the resolution difference of each $\mathbf{F}_r^{(i)}$, the dilation rates of each unit are different and the details are presented in Tab. II. In this unit, the dilated convolutions large the receptive field without increasing the computation. They are performed in a parallel manner, which makes $\mathbf{F}_{mi}^{(i)}$ effectively capture the local and global location information of the gazed objects.

The multi-scale features in $\mathbf{F}_{mi}^{(i)}$ are complementary to each other. They are blended to produce the location response maps $\mathbf{r}_{loc}^{(i)} \in [0, 1]^{h_i \times w_i \times c_i}$ via:

$$\mathbf{F}_{int}^{(i)} = 2C(\mathbf{F}_{mi}^{(i)}; W_{2c}^{(i)}), \tag{4}$$

$$\mathbf{r}_{loc}^{(i)} = \psi(C(\mathbf{F}_{int}^{(i)}; W_c^{(i)})), \tag{5}$$

where $\mathbf{F}_{int}^{(i)}$ is the interim feature, $2C(*; W_{2c}^{(i)})$ are two convolutional layers with the same parameters $W_{2c}^{(i)}$, $\psi(\cdot)$ is the sigmoid function, and $C(*; W_c^{(i)})$ is the convolutional layer with parameters $W_c^{(i)}$ which are 3×3 kernel with c_i channels. $W_{2c}^{(i)}$ contain kernel size and channel number, which are different in different OLMs. Their details are shown in the column with "2×Conv" of Tab. II.

TABLE II

Detailed Parameters of Each OLM. We Present the Kernel Size and Channel Number of Each Dilated/Normal Convolutions. Besides, We Also Present the Dilation Rates and the Size of Output Feature. For Instance, $(3 \times 3, 32)$ Denotes That the Kernel Size Is 3×3 and the Channel Number Is 32

Aspects	Dilation conv	Dilation rate	2×Conv	Output size
OLM-1	$(3 \times 3, 32)$	1/3/5/7	$(7 \times 7, 128)$	$[288 \times 288 \times 128]$
OLM-2	$(3 \times 3, 64)$	1/3/5/7	$(5 \times 5, 256)$	$[144 \times 144 \times 256]$
OLM-3	$(3 \times 3, 128)$	1/3/5/7	$(5 \times 5, 512)$	$[72 \times 72 \times 512]$
OLM-4	$(3 \times 3, 256)$	1/2/3/4	$(3 \times 3, 1024)$	$[36 \times 36 \times 1024]$
OLM-5	$(3 \times 3, 256)$	1/2/3/4	$(3 \times 3, 1024)$	$[18 \times 18 \times 1024]$

After completing the FDM interpretation in location analysis unit, we successfully obtain $\mathbf{r}_{loc}^{(i)}$, which are the protagonists of the feature re-weighting (*i.e.* Re-wei) part. We employ $\mathbf{r}_{loc}^{(i)}$ to re-weight $\mathbf{F}_r^{(i)}$ at channel-wise and spatial-wise, and receive the location-enhanced feature $\mathbf{F}_{loc}^{(i)} \in \mathbb{R}^{h_i \times w_i \times c_i}$, which is computed as:

$$\mathbf{F}_{loc}^{(i)} = \mathbf{F}_r^{(i)} \otimes \mathbf{r}_{loc}^{(i)}, \tag{6}$$

where $\otimes$ is element-wise multiplication. Besides, in Re-wei, to balance the information of image and location, we concatenate $\mathbf{F}_r^{(i)}$ to $\mathbf{F}_{loc}^{(i)}$ and obtain the output feature $\mathbf{F}_{olm}^{(i)}$ of OLM. The size of $\mathbf{F}_{olm}^{(i)}$ is shown in Tab. II. Notably,

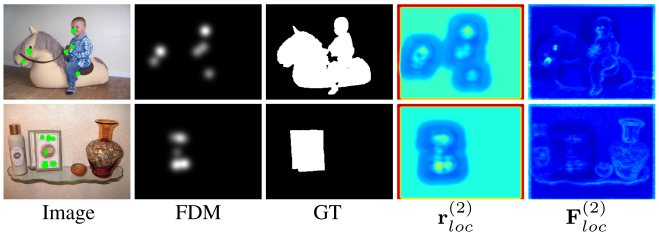


Fig. 4. Feature visualization in OLM-2. $\mathbf{r}_{loc}^{(2)}$ is the location response map, and $\mathbf{F}_{loc}^{(2)}$ is the location-enhanced feature.

at the training phase, we apply the pixel-level segmentation supervision (*i.e.* Seg sup) to each OLM.

In Fig. 4, we visualize feature in OLM-2 to verify the effectiveness of the location enhancement. Concretely, in OLM-2, the *conv2-2* is re-weighted by the location response map. As shown in Fig. 4, the location response map $\mathbf{r}_{loc}^{(2)}$ contains rich location information of the gazed objects. After using Eq. 6 to perform the location enhancement operation on *conv2-2*, we observe that the gazed objects are highlighted in $\mathbf{F}_{loc}^{(2)}$ (with darker color). In summary, the location-enhanced feature $\mathbf{F}_{loc}^{(i)}$ of OLM has strong location expression ability and contributes to the subsequent segmentation prediction network.

D. Boundary Preservation Module

The *Boundary Preservation Module* is built to restrain the falsely highlighted part of the re-weighted feature of OLM and to preserve the completeness of the gazed objects for the segmentation prediction. As the **BPM-5** shown in Fig. 3, the structure of BPM is succinct, but it is a key bridge to connect convolutional blocks of the prediction network.

Let $\{\mathbf{F}_p^{(i)}: \mathbf{F}_p^{(i)} \in \mathbb{R}^{h_{i-1} \times w_{i-1} \times c_{i-1}}, i = 2, 3, 4, 5\}$ denote the output feature of each deconvolutional layer in the prediction network. In BMP, $\mathbf{F}_p^{(i)}$ is processed by a convolutional layer to generate the boundary mask $\mathbf{B}^{(i)}$, which is defined as:

$$\mathbf{B}^{(i)} = C(\mathbf{F}_p^{(i)}, W_c^{(i)}). \tag{7}$$

To increase the accuracy of $\mathbf{B}^{(i)}{}_{i \in \{2,3,4,5\}}$, we introduce the pixel-level boundary supervision (*i.e.* "Bound sup" on **BPM-5** in Fig. 3) in BPM. Since that there are no pixel-level boundary annotations in the PFOS dataset, we employ the morphological operation on binary segmentation GT $\mathbf{G}_s$ to produce the boundary GT $\mathbf{G}_b$, as follow:

$$\mathbf{G}_b = \mathrm{Dilate}(\mathbf{G}_s; \theta) - \mathbf{G}_s, \tag{8}$$

where $\mathrm{Dilate}(*; \theta)$ is the morphological dilation operation with dilation coefficient θ which is 2 pixels.

Then, $\mathbf{B}^{(i)}$ is concatenated to $\mathbf{F}_p^{(i)}$ to generate the output feature $\mathbf{F}_{bpm}^{(i)}$ of BPM. We also put the pixel-level segmentation supervision behind $\mathbf{F}_{bpm}^{(i)}$, such as "Seg sup" on **BPM-5** in Fig. 3. The segmentation supervision and the boundary supervision cooperate well with each other, improving the feature representation of the gazed objects. In this way, we novelly introduce boundary information into the BPM, and $\mathbf{F}_{bpm}^{(i)}$ carries the feature de-noising and boundary preservation capabilities into the prediction network.

E. Implementation Details

1) Prediction Network: The prediction network is constructed in the top-down manner to gradually restore resolution. It consists of five convolutional blocks, four BPMs and four deconvolutional layers. A dropout layer [61] is placed before each deconvolutional layer to prevent the prediction network from overfitting. In addition, we attach the boundary prediction branch to the prediction network to assist the object segmentation branch. We initialize parameters of the prediction network by xavier method [62].

2) Overall Loss: As shown in Fig. 3, there are totally 15 losses in the OLBP network, including 10 segmentation losses and 5 boundary losses. The overall loss $\mathbb{L}$ can be divided into three parts: losses of multi-task prediction, losses on OLMs and losses on BPMs. $\mathbb{L}$ is calculated as:

$$\begin{aligned}\mathbb{L} = [\mathcal{L}_s(\mathbf{S}^{(1)}, \mathbf{G}_s) + \mathcal{L}_s(\mathbf{B}^{(1)}, \mathbf{G}_b)] + \sum_{i=1}^{5} \mathcal{L}_s(\mathbf{S}_{olm}^{(i)}, \mathbf{G}_s) \\ + \sum_{i=2}^{5} [\mathcal{L}_s(\mathbf{S}_{bpm}^{(i)}, \mathbf{G}_s) + \mathcal{L}_s(\mathbf{B}^{(i)}, \mathbf{G}_b)],\end{aligned} \tag{9}$$

where $\mathcal{L}_s(\cdot, \cdot)$ is the softmax loss, $\mathbf{S}^{(1)}$ is the predicted segmentation map, and $\mathbf{B}^{(1)}$ is the predicted boundary map. $\mathbf{S}_{olm}^{(i)}$ and $\mathbf{S}_{bpm}^{(i)}$ present the side output segmentation results in OLM and BPM, respectively. $\mathbf{B}^{(i)}{}_{i \in \{2,3,4,5\}}$ is the boundary mask in BPM. Notably, for each softmax loss, we resize the resolutions of $\mathbf{G}_s$ and $\mathbf{G}_b$ to fit the resolutions of corresponding $\mathbf{S}_{olm}^{(i)}$, $\mathbf{S}_{bpm}^{(i)}$ and $\mathbf{B}^{(i)}$.

3) Network Training: The PFOS dataset is separated into training set and testing set. The training set contains 600 images with 9,000 personal fixation maps, including 3,075 constrained fixation maps and 5,925 unconstrained fixation maps. The testing set consists of 100 images with 1,500 personal fixations, including 608 constrained fixation maps and 892 unconstrained fixation maps.

The OLBP network is implemented on Caffe [63] and experimented using a NVIDIA Titan X GPU. The data of training set and testing set are resized to 288 × 288 for training and inference. We adopt the standard stochastic gradient descent (SGD) method [64] to optimize our OLBP network for 30,000 iterations. The learning rate is set to 8×10^{-8}, and it will be divided by 10 after 14,000 iterations. The dropout ratio, batch size, iteration size, momentum and weight decay are set to 0.5, 1, 8, 0.9 and 0.0001, respectively.

VI. Experiments

In this section, we present comprehensive experiments on the proposed PFOS dataset. We introduce evaluation metrics in Sec. VI-A. In Sec. VI-B, we compare the proposed OLBP network with state-of-the-art methods. Then, we conduct ablation studies in Sec. VI-C and show some personal segmentation results in Sec. VI-D. Finally, we present some discussions on the connections between fixation-based object segmentation and salient object detection in Sec. VI-E.

A. Evaluation Metrics

We use five evaluation metrics, *i.e*. Jaccard index ($\mathcal{J}$), S-measure ($\mathcal{S}_\lambda$) [65], F-measure ($\mathcal{F}_\beta$), weighted F-measure ($w\mathcal{F}_\beta$) [66], and E-measure ($\mathcal{E}_\xi$) [67], to evaluate the performance of different methods.

1) Jaccard Index $\mathcal{J}$: Jaccard index is also called intersection-over-union (IoU), which can compare similarities and differences between two binary maps. It is defined as:

$$\mathcal{J} = \frac{|\mathbf{S} \cap \mathbf{G}_s|}{|\mathbf{S} \cup \mathbf{G}_s|}, \tag{10}$$

where $\mathbf{S}$ is the predicted segmentation map, and $\mathbf{G}_s$ is the binary segmentation GT.

2) S-Measure $\mathcal{S}_\lambda$: S-measure focuses on the structural similarity between the predicted segmentation map and the binary segmentation GT. It evaluates the structural similarity of region-aware (S_r) and object-aware (S_o) simultaneously. S-measure is defined as:

$$\mathcal{S}_\lambda = \lambda * S_o + (1 - \lambda) * S_r, \tag{11}$$

where λ is set to 0.5 by default.

3) F-Measure $\mathcal{F}_\beta$: F-measure is a weighted harmonic mean of precision and recall, which considers precision and recall comprehensively. It is defined as:

$$\mathcal{F}_\beta = \frac{(1+\beta^2) \times Precision \times Recall}{\beta^2 \times Precision + Recall}, \tag{12}$$

where β^2 is set to 0.3 following previous studies [47], [48].

4) Weighted F-Measure $w\mathcal{F}_\beta$: Weighted F-measure has the ability to evaluate the non-binary and binary map. It focuses on evaluating the weights errors of predicted pixels according to their location and their neighborhood, which is formulated as:

$$w\mathcal{F}_\beta = \frac{(1+\beta^2) \times Precision^w \times Recall^w}{\beta^2 \times Precision^w + Recall^w}, \tag{13}$$

where β^2 is set to 1 following previous studies [68], [69].

5) E-Measure $\mathcal{E}_\xi$: E-measure is based on cognitive vision studies. It evaluates the local errors (*i.e*. pixel-level) and the global errors (*i.e*. image-level) together. We introduce it to provide a more comprehensive evaluation. It could be computed as:

$$\mathcal{E}_\xi = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} f\Big(\frac{2\varphi_{\mathbf{G}_s} \circ \varphi_{\mathbf{S}}}{\varphi_{\mathbf{G}_s} \circ \varphi_{\mathbf{G}_s} + \varphi_{\mathbf{S}} \circ \varphi_{\mathbf{S}}}\Big), \tag{14}$$

where $\varphi_{\mathbf{G}_s}$ and $\varphi_{\mathbf{S}}$ are distance bias matrices for binary segmentation GT and predicted segmentation map, respectively, $\circ$ is the Hadamard product, and $f(\cdot)$ is the quadratic form.

B. Comparison With the State-of-the-Arts

1) Comparison Methods: We compare our OLBP network against three types of state-of-the-art methods, including *semantic segmentation-based methods*, *clicks-based interactive image segmentation methods* and *fixations-based object segmentation methods*. For a reasonable comparison of the first type of method, we follow [12], [25], which convert the segmentation problem into the selection problem. Concretely, we first apply semantic segmentation methods, *i.e*. PSPNet [54], SegNet [55], DeepLab [51], EncNet [56], DeepLabV3+ [57], and HRNetV2 [58], to image, and then use the fixations to select the gazed objects. The second type of method includes ISLD [14], FCTSFN [17], and BRS [16]. The last type of method includes AVS [22], SOS [23], GBOS [24] and CFPS [25]. For all the above compared methods, we use the implementations with recommend parameter settings for a fair comparison.

In addition, we modify several semantic segmentation methods (*i.e*. DeepLabV3+ [57] and HRNetV2 [58]) and recent salient object detection methods (*i.e*. CPD [59] and GCPA [60]) by embedding FDM in them to guide object segmentation. Two types of comparison methods are thus generated, namely *FDM-guided semantic segmentation* and *FDM-guided salient object detection*, respectively. Specifically, for DeepLabV3+, we embed FDM into features (*i.e*. low-level features and features generated from the ASPP) to bridge the encoder and decoder; for HRNetV2, we embed FDM between the second stage and the third stage; for CPD, we embed FDM into two partial decoders; and, for GCPA, we embed FDM into four self refinement modules. We retrain these modified methods with the same training dataset as our method, and their parameters are adjusted for better convergence. Notably, we use the well-known OTSU method [70] to binarize the generated probability map of our method and other CNNs-based methods.

2) Quantitative Performance Evaluation: We evaluate our OLBP network and other 17 state-of-the-art methods on the PFOS dataset using above five evaluation metrics. The quantitative results are presented in Table III. Our OLBP network favorably outperforms all the compared methods in terms of different metrics. Concretely, compared with the best method CFPS [25] in fixations-based object segmentation methods, the performance of our method is improved by 3.2%, 2.2% and 3.0% in $\mathcal{J}$, $\mathcal{S}_\lambda$ and $w\mathcal{F}_\beta$, respectively. The performance of our method is 5.9% better than FCTSFN [17] in $\mathcal{E}_\xi$, and is 6.4% better than ISLD [14] in $\mathcal{F}_\beta$. Note that the performance of our method is far better than that of three traditional methods AVS [22], SOS [24] and GBOS [24]. We attribute the performance superiority of the proposed OLBP network to the scheme of object localization and boundary preservation.

In addition, semantic segmentation-based methods get an average of 51.6% in $\mathcal{J}$. This may be due to the fact that semantic segmentation methods cannot accurately segment all objects, resulting in the failure of the object selection process. Clicks-based interactive image segmentation methods achieve an average of 61.9% in $\mathcal{J}$, while our OLBP network obtains 73.7% in $\mathcal{J}$. This demonstrates that our method is more robust than clicks-based interactive image segmentation methods in adapting the ambiguity of fixations. Fixations-based object segmentation methods contain three traditional methods and one CNN-based method, obtaining an average of 48.0% in $\mathcal{J}$.

Specifically, we present the results of the FDM-guided semantic segmentation methods, including the modified DeepLabV3+ and HRNetV2, in Table III. The modified

TABLE III

Quantitative Results Including Jaccard Index, S-Measure, Weighted F-Measure, E-Measure and F-Measure on the PFOS Dataset (in Percentage %). *Semantic Segmentation* Means Semantic Segmentation-Based Method. *Clicks* Means Clicks-Based Interactive Image Segmentation Method. *Fixations* Means Fixations-Based Object Segmentation Method. *FDM-Guided Semantic Segmentation* Means Embedding FDM Into Semantic Segmentation Method. *FDM-Guided Salient Object Detection* Means Embedding FDM Into Salient Object Detection Method. The Best Three Results Are Shown in Red, Blue, and Green. ↑ Denotes Larger Is Better. The Subscript of Each Method Represents the Publication Year. † Means CNNs-Based Method

Aspects	Methods	PFOS Dataset				
		$\mathcal{J}\uparrow$	$\mathcal{S}_\lambda\uparrow$	$w\mathcal{F}_\beta\uparrow$	$\mathcal{E}_\xi\uparrow$	$\mathcal{F}_\beta\uparrow$
Semantic Segmentation	PSPNet$_{17}^{\dagger}$ [54]	51.0	58.9	55.5	64.2	60.2
	SegNet$_{17}^{\dagger}$ [55]	58.7	70.4	66.6	78.4	72.5
	DeepLab$_{18}^{\dagger}$ [51]	52.8	65.7	60.5	72.9	66.9
	EncNet$_{18}^{\dagger}$ [56]	55.5	62.2	60.5	69.0	65.3
	DeepLabV3+$_{18}^{\dagger}$ [57]	45.6	61.4	53.1	67.8	59.3
	HRNetV2$_{19}^{\dagger}$ [58]	46.1	50.7	49.0	53.8	53.2
Clicks	ISLD$_{18}^{\dagger}$ [14]	61.2	73.4	71.2	82.5	77.9
	FCTSFN$_{19}^{\dagger}$ [17]	62.4	72.9	69.9	82.8	75.1
	BRS$_{19}^{\dagger}$ [16]	62.1	73.0	69.1	82.3	74.6
Fixations	AVS$_{12}$ [22]	40.9	56.0	48.7	65.1	56.6
	SOS$_{14}$ [23]	42.6	57.5	51.4	67.8	60.0
	GBOS$_{17}$ [24]	38.0	56.7	48.0	63.9	58.1
	CFPS$_{19}^{\dagger}$ [25]	70.5	78.9	76.7	87.4	81.3
FDM-Guided Semantic Segmentation	DeepLabV3+$_{18}^{\dagger}$ [57]	**71.0**	**79.5**	**78.3**	**87.6**	**83.2**
	HRNetV2$_{19}^{\dagger}$ [58]	58.8	71.3	68.6	80.4	75.7
FDM-Guided Salient Object Detection	CPD$_{19}^{\dagger}$ [59]	69.2	78.4	76.4	86.2	81.7
	GCPA$_{20}^{\dagger}$ [60]	**72.3**	**80.3**	**78.9**	**88.1**	**83.6**
Personal Fixations	**OLBP (Ours)**	**73.7**	**81.1**	**80.0**	**88.7**	**84.3**

DeepLabV3+ achieves a promising performance, but does not exceed our OLBP network (*e.g.* 71.0% *vs* 73.7% in $\mathcal{J}$). Although the FDM guidance brings some advantages to HRNetV2, but the modified HRNetV2 still does not perform well. For the FDM-guided salient object detection, both modified CPD and GCPA perform well, though our OLBP still outperforms them (*e.g.* 4.5% and 1.4% better than the modified CPD and GCPA in $\mathcal{J}$, respectively). In summary, there is a large room for performance improvement on the proposed PFOS dataset, suggesting that the PFOS dataset is challenging to all compared methods including OLBP.

3) Qualitative Performance Evaluation: In Fig. 5, we show some representative visualization results of our OLBP network and other methods. Obviously, the visual segmentation maps of three traditional methods GBOS [24], SOS [24] and AVS [22] are rough. However, the CNN-based method CFPS [25], which belongs to the same type as GBOS, SOS and AVS, basically captures the gazed objects and brings in less background regions. The gazed objects in the segmentation results of clicks-based interactive image segmentation methods BRS [16], FCTSFN [17], and ISLD [14] are partially segmented and the details are relatively coarse. As for the EncNet [56], DeepLab [51] and SegNet [55], the object segmentation maps of them depend on the semantic segmentation results, which are great uncertainty. This results in their object segmentation maps that are sometimes accurate and sometimes bad.

In contrast, our OLBP network is equipped with the scheme of object localization and boundary preservation, which precisely analyzes the location information of fixations and completes the gazed objects. The segmentation maps of "Ours" in Fig. 5 are very localized in the gazed objects with pretty fine details, even under the interference of some ambiguous fixations.

4) Robustness Evaluation: We provide a robustness evaluation of our method and several representative methods, including the modified GCPA [60], CFPS [25] and the modified CPD [59], on the test dataset of the PFOS dataset. Concretely, we add the noise, *i.e.* unconstrained fixations, to the fixation map by random sampling on the background regions at three levels, *i.e.* different percentages (15%, 30%, 45%) increase in the number of unconstrained fixations of the total number of fixations. The performance of above methods after adding noise are presented in Table IV. Our method consistently outperforms the compared methods under three challenging situations, showing excellent robustness.

C. Ablation Studies

We comprehensively evaluate the contribution of each vital component to performance in our OLBP network. Specifically, we assess 1) the overall contributions of OLM and BPM; 2) the effectiveness of the three parts in OLM; and 3) the usefulness of BPM and the top-down manner in prediction network. The variants are retrained with the same hyper-parameters and

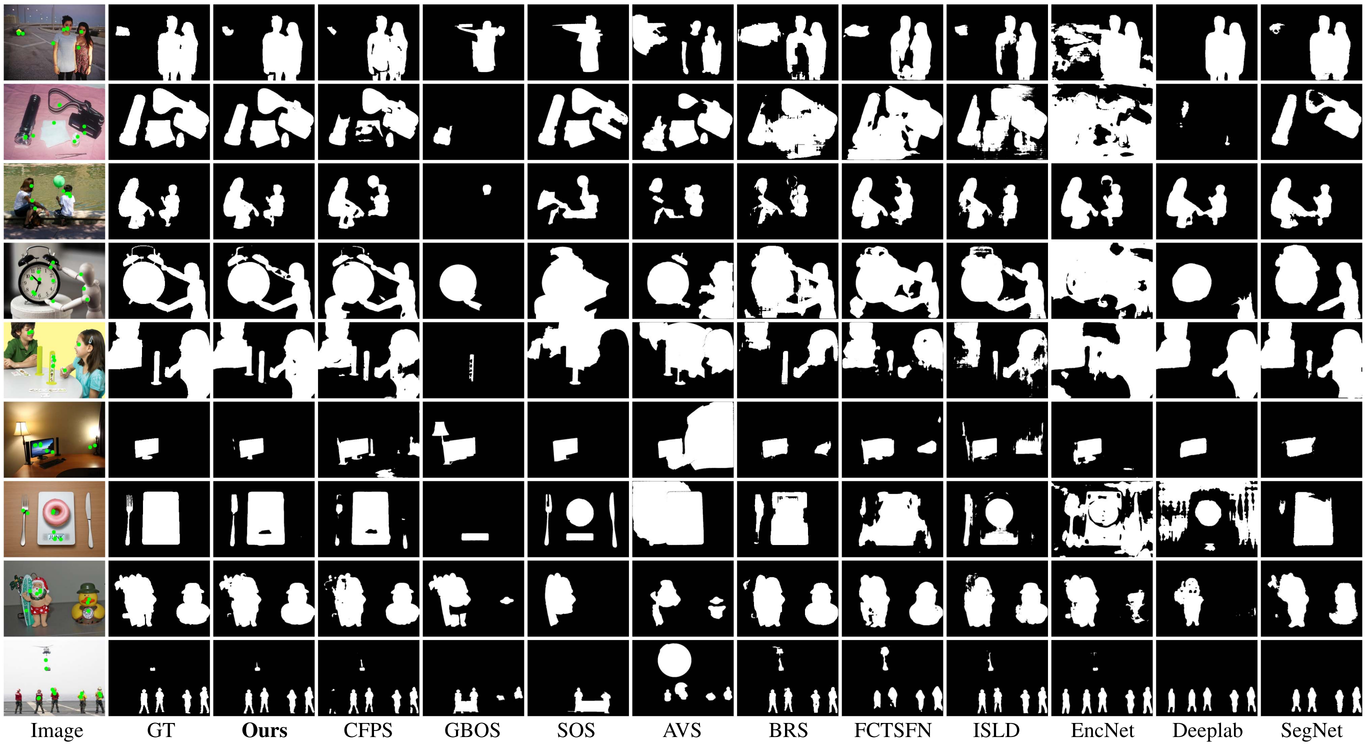


Fig. 5. Visualization comparison to some representative methods on the PFOS dataset. Zoom-in for the best view.

TABLE IV

Robustness Evaluation of Our Method and Several Representative Methods, Such as the Modified GCPA [60], CFPS [25] and the Modified CPD [59], on the Test Part of the PFOS Dataset in Terms of Jaccard Index. The Best Result of Each Row Is Shown in **Bold**. Notably, "+15% Noise" Means an Additional 15% Increase in the Number of Unconstrained Fixations of the Total Number of Fixations in a Fixation Map. We Add the Noise (*i.e.* Unconstrained Fixations) at Three Levels, *i.e.* 15%, 30%, and 45%

Dataset	**OLBP (Ours)**	GCPA_{20} [60]	CFPS_{19} [25]	CPD_{19} [59]
PFOS	**73.7**	72.3	70.5	69.2
+15% noise	**72.2**	70.9	69.6	68.7
+30% noise	**71.3**	70.1	69.2	68.4
+45% noise	**70.3**	69.7	68.8	68.1

TABLE V

Ablation Analyses for the Proposed OLBP Network on the PFOS Dataset (in Percentage %). As Can Be Observed, Each Component in OLBP Network Plays an Important Role and Contributes to the Performance. The Best Result in Each Column Is **Bold**. Baseline: Encoder-Decoder Network, OLM: Object Localization Module, and BPM: Boundary Preservation Module

	Baseline	OLM	BPM	$\mathcal{J}\uparrow$	$\mathcal{S}_\lambda\uparrow$	$w\mathcal{F}_\beta\uparrow$
1	✓*			67.2	75.9	72.2
2	✓*		✓	68.0	76.4	72.5
3	✓			70.7	78.3	75.0
4	✓	✓		73.0	80.7	79.5
5	✓		✓	71.4	78.7	75.6
6	✓	✓	✓	**73.7**	**81.1**	**80.0**

✓* means the image and FDM are concatenated.

✓means the image and FDM are fed to network separately.

training set as aforementioned settings in Sec. V-E, and the experiments are conducted on the PFOS dataset.

1. Does the proposed OLM and BPM contribute to OLBP network? To evaluate the contribution of the proposed OLM and BPM to OLBP network, we derive three variants: baseline network (denoted by "Ba"/"Ba*"), baseline network with only OLMs ("Ba+OLM"), and baseline network with only BPMs ("Ba/Ba* + BPM"). In particular, we provide two types of baseline network: the first one is an encoder-decoder network, whose input is the concatenated image and FDM (denoted by "Ba*"); the second one is an encoder-decoder network with the down-sampled FDMs being concatenated to each skip-layer (denoted by "Ba"), *i.e.* the image and FDM are fed to network separately. We report the quantitative results in Tab. V.

We observe that the first baseline network "Ba*" (the 1st line in Tab. V) only obtains 67.2% in $\mathcal{J}$, and the second baseline network "Ba" (the 3rd line in Tab. V) obtains 70.7% in $\mathcal{J}$. This confirms that direct concatenation of the image and FDM results in the location information of FDM being submerged by image information; by contrast, concatenating FDM with image features at each scale benefits object location. OLM significantly improves the performance of the

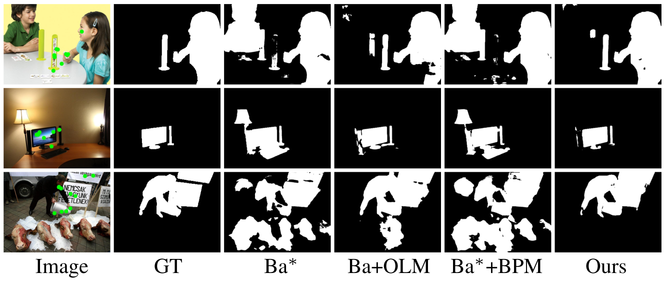


Fig. 6. Visual comparisons of different variants. “Ba*” is the baseline network, whose input is the concatenated image and FDM.

baseline network (*e.g.* $\mathcal{J}: 67.2\%/70.7\% \rightarrow 73.0\%$ and $w\mathcal{F}_\beta$: $72.2\%/75.0\% \rightarrow 79.5\%$). This shows that the contribution of OLM is remarkable, and OLM does capture the location information. Comparing with OLM, the contribution of BPM to baseline networks is slightly inferior (*e.g.* $\mathcal{J}: 67.2\% \rightarrow 68.0\%; 70.7\% \rightarrow 71.4\%$), but BPM also shows its effectiveness to improve performance of “Ba+OLM” (*e.g.* $w\mathcal{F}_\beta: 79.5\% \rightarrow 80.0\%$). This demonstrates that BPM can further complete the objects and filter background of erroneous localization. With the cooperation between OLM and BPM, the performance of the whole OLBP network is improved by 6.5%/3.0% in $\mathcal{J}$, 5.2%/2.8% in $\mathcal{S}_\lambda$ and 7.8%/5.0% in $w\mathcal{F}_\beta$ compared with the baseline network “Ba*”/“Ba”. This demonstrates that the scheme of bottom-up object localization and top-down boundary preservation is successfully embedded into the baseline network.

Additionally, the segmentation maps of variants based on the first baseline network “Ba*” are shown in Fig. 6. We observe that “Ba*” almost segments all the objects in images. With the assistance of OLM, “Ba* + OLM” determines the location of the gazed objects, and the gazed objects on the segmentation maps of “Ba+OLM” are much clearer. Finally, with the help of BPM, the segmentation maps of ours (*i.e.* OLBP network) are satisfactory.

2. How effective are the three parts in OLM? As described in Sec. V-C, OLM consists of location analysis unit, feature re-weighting (*i.e.* Re-wei) and segmentation supervision (*i.e.* Seg sup). To validate the effectiveness of the three parts in OLM, we modify the structure of OLM and provide four variants: a) the four dilated convolutions are replaced by one convolutional layer in the location analysis unit (*w/o dilated convs*); b) without using response maps to re-weight image feature in Re-wei (*w/o multiply*); c) without concatenating re-weighted feature and image feature in Re-wei (*w/o concat*); and d) without segmentation supervision (*w/o Seg sup*). The ablation results are reported in Tab. VI, and the detailed structures of the above four OLM variants are presented in Fig. 7.

We discover that the performances of the four variants are worse than ours. Concretely, the performance degradation of *w/o dilated convs* (*e.g.* $\mathcal{J}: 73.7\% \rightarrow 72.7\%$) validates that the parallel dilated convolutions analyze FDM thoroughly and one convolutional layer cannot mine sufficient location information from FDM. The performance drop of *w/o multiply* (*e.g.* $\mathcal{S}_\lambda: 81.1\% \rightarrow 80.5\%$) confirms that the location response maps are more suitable to highlight objects on CNN feature of image than using them directly. The reason behind this is that location response maps are a group of probability maps, without rich object, texture and color information. Besides, *w/o concat* brings 0.9% performance penalty in $\mathcal{J}$, which shows that the information balance between image and location is important. *w/o Seg sup* carries 0.6% performance drop in $w\mathcal{F}_\beta$. This demonstrates that the segmentation supervision can enhance representation of the gazed objects.

TABLE VI

ABLATION RESULTS OF THE OLM ON THE PFOS DATASET (IN PERCENTAGE %). THE BEST RESULT IN EACH COLUMN IS **BOLD**. THE CORRESPONDING STRUCTURES OF THE LISTED VARIANTS ARE PRESENTED IN FIG. 7

OLM variants	$\mathcal{J}\uparrow$	$\mathcal{S}_\lambda\uparrow$	$w\mathcal{F}_\beta\uparrow$
w/o dilated convs	72.7 -1.0	80.7 -0.4	79.0 -1.0
w/o multiply	72.6 -1.1	80.5 -0.6	79.2 -0.8
w/o concat	72.8 -0.9	80.8 -0.3	79.6 -0.4
w/o Seg sup	72.9 -0.8	80.9 -0.2	79.4 -0.6
Ours	**73.7**	**81.1**	**80.0**

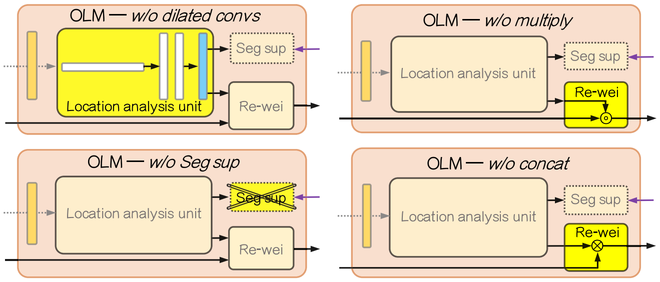


Fig. 7. Structures of four OLM variants. *w/o dilated convs*: the four dilated convolutions are replaced by one convolutional layer; *w/o multiply*: without using response maps to re-weight image feature in Re-wei; *w/o concat*: without concatenating re-weighted feature and image feature in Re-wei; *w/o Seg sup*: without segmentation supervision.

3. Is it useful to adopt BPM and the top-down manner in prediction network? To investigate the usefulness of the top-down manner in prediction network, we report the performance of side output segmentation maps of BPM in Tab. VII. Besides, we also report the side output performance of *w/o BPM* in Tab. VII to evaluate the importance of BPM.

We observe that the quantitative results of side outputs ($\mathbf{S}^{(5)}_{bpm}$, $\mathbf{S}^{(4)}_{bpm}$, $\mathbf{S}^{(3)}_{bpm}$, $\mathbf{S}^{(2)}_{bpm}$ and $\mathbf{S}^{(1)}$) are incremental in terms of both *w/ BPM* (*e.g.* $w\mathcal{F}_\beta: 67.8\% \rightarrow 75.5\% \rightarrow 78.9\% \rightarrow 79.9\% \rightarrow 80.0\%$) and *w/o BPM* (*e.g.* $\mathcal{S}_\lambda: 70.4\% \rightarrow 76.9\% \rightarrow 79.8\% \rightarrow 80.6\% \rightarrow 80.7\%$). This confirms that the top-down manner is useful for the prediction network. The differences between the performance of *w/o BPM* and *w/ BPM* are also reported in Tab. VII. We discover that all the differences are negative, which shows that BPM works well for each side output of the top-down prediction network.

D. Personal Segmentation Results

Due that the personal fixations are closely related to age and gender, different users are interested in different objects when

TABLE VII

THE PERFORMANCE OF SIDE OUTPUT SEGMENTATION MAPS OF WITH/WITHOUT BPM ON PFOS DATASET (IN PERCENTAGE %). THE NUMBER IN THE LOWER RIGHT CORNER OF THE PERFORMANCE OF W/O BPM IS THE DIFFERENCE BETWEEN IT AND THE PERFORMANCE OF W/ BPM. THE BEST RESULT IN EACH COLUMN IS **BOLD**

Side outputs	*w/ BPM* (**Ours**)			*w/o BPM*		
	$\mathcal{J}\uparrow$	$\mathcal{S}_\lambda\uparrow$	$w\mathcal{F}_\beta\uparrow$	$\mathcal{J}\uparrow$	$\mathcal{S}_\lambda\uparrow$	$w\mathcal{F}_\beta\uparrow$
$\mathbf{S}_{bpm}^{(5)}$	62.0	71.6	67.8	60.2 $_{-1.8}$	70.4 $_{-1.2}$	66.0 $_{-1.8}$
$\mathbf{S}_{bpm}^{(4)}$	69.2	77.5	75.5	68.2 $_{-1.0}$	76.9 $_{-0.6}$	74.8 $_{-0.7}$
$\mathbf{S}_{bpm}^{(3)}$	72.6	80.2	78.9	71.9 $_{-0.7}$	79.8 $_{-0.4}$	78.3 $_{-0.6}$
$\mathbf{S}_{bpm}^{(2)}$	73.7	81.1	79.9	72.9 $_{-0.8}$	80.6 $_{-0.5}$	79.4 $_{-0.5}$
$\mathbf{S}^{(1)}$	**73.7**	**81.1**	**80.0**	**73.0** $_{-0.7}$	**80.7** $_{-0.4}$	**79.5** $_{-0.5}$

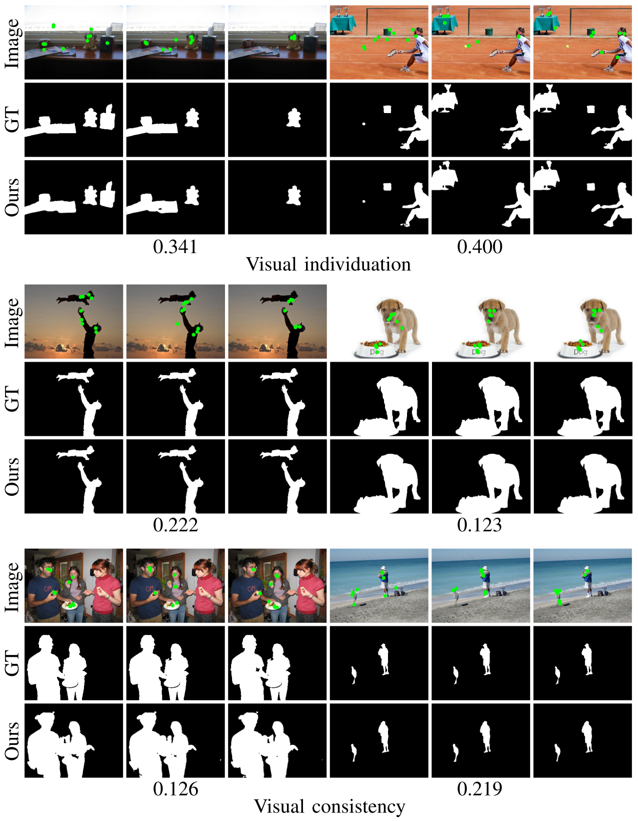


Fig. 8. Visual examples of personal segmentation results. There are two basic properties of personal visual systems: visual individuation and visual consistency. The value of each image is the mean JS score.

observing the same scene. We define the visual difference of different personal visual systems as visual individuation. Some examples of visual individuation are presented in the first part of Fig. 8. We can observe that there are multiple different types of objects and complex backgrounds in these images. The personal fixations of different users are located on different objects, which correspond to the distinctive GTs.

In addition, we discover that personal visual systems are also consistent in some scenes, which is denoted as visual consistency. We show some examples of visual consistency in the second and third parts of Fig. 8. The images in the second part contain simple backgrounds and sparse objects, and the images in the third part contain more competitive situation, *i.e.* complex background and partially selected objects. In both parts, we observe that the locations of different personal fixations are similar, resulting in the identical GTs of different users. Notably, in either case, our method show the ability to segment the gazed objects consistent with the corresponding GT.

We also provide the quantitative analysis of visual individuation and visual consistency with Jensen-Shannon (JS) divergence. JS divergence evaluates the similarity of two probability distributions $\mathbf{S}^1$ and $\mathbf{S}^2$, and it is based on Kullback-Leibler (KL) divergence. Its value belongs to [0, 1]. The closer its value is to zero, the smaller the difference between $\mathbf{S}^1$ and $\mathbf{S}^2$ is and the more similar they are. It can be expressed as follows:

$$\mathrm{JS}(\mathbf{S}^1,\mathbf{S}^2)=\frac{1}{2}\mathrm{KL}(\mathbf{S}^1,\frac{\mathbf{S}^1+\mathbf{S}^2}{2})+\frac{1}{2}\mathrm{KL}(\mathbf{S}^2,\frac{\mathbf{S}^1+\mathbf{S}^2}{2}), \tag{15}$$

$$\mathrm{KL}(\mathbf{P},\mathbf{Q})=\sum_{i=1}^{N}\mathbf{P}_i\log\left(\epsilon+\frac{\mathbf{P}_i}{\epsilon+\mathbf{Q}_i}\right), \tag{16}$$

where KL(·) is Kullback-Leibler divergence, which is often used as an evaluation metric in fixation prediction [71]–[74], i indicates the i^{th} pixel in the probability distribution, N is the total number of pixels, and ϵ is a regularization constant.

We introduce JS to measure the similarity of fixation points maps of each image in Fig. 8. First, we transform the fixation points map (green dots in each image) to FDM using Eq. 1; then we compute the JS score of each two FDMs; finally we report the mean JS score for each image in Fig. 8. It is obvious that the mean JS scores (*i.e.* 0.222, 0.123, 0.126, and 0.219) of images which belong to visual consistency are relatively smaller than those (*i.e.* 0.341 and 0.400) of images which belong to visual individuation. And the mean JS scores of images which belong to visual consistency are close to zero, which indicates that the distributions of FDMs are very similar, *i.e.* people may look at the same object(s).

E. Discussions

Salient Object Detection (SOD) is widely explored in color images [59], [60], [75]–[77], RGB-D images [78], [79] and videos [80]–[82], and it is closely related to our fixation-based object segmentation task. In this section, we discuss the connections between fixation-based object segmentation and SOD.

SOD aims to highlight the most visually attractive object(s) in a scene, while fixation-based object segmentation aims to segment the gazed objects according to the fixation map, as defined in Sec. III. To illustrate the differences and connections between these two tasks, we conduct experiments on two SOD datasets, *i.e.* DUTS-OMRON [46] and PASCAL-S [23], and show visual comparisons with two state-of-the-art SOD methods, *i.e.* CPD [59] and GCPA [60], in Fig. 9, which summarizes three situations. First, in the 1st and 2nd rows, we present the differences of these two tasks: our method not only segments the salient objects, such as the bird and the big tent, but also segments the gazed wood stake and

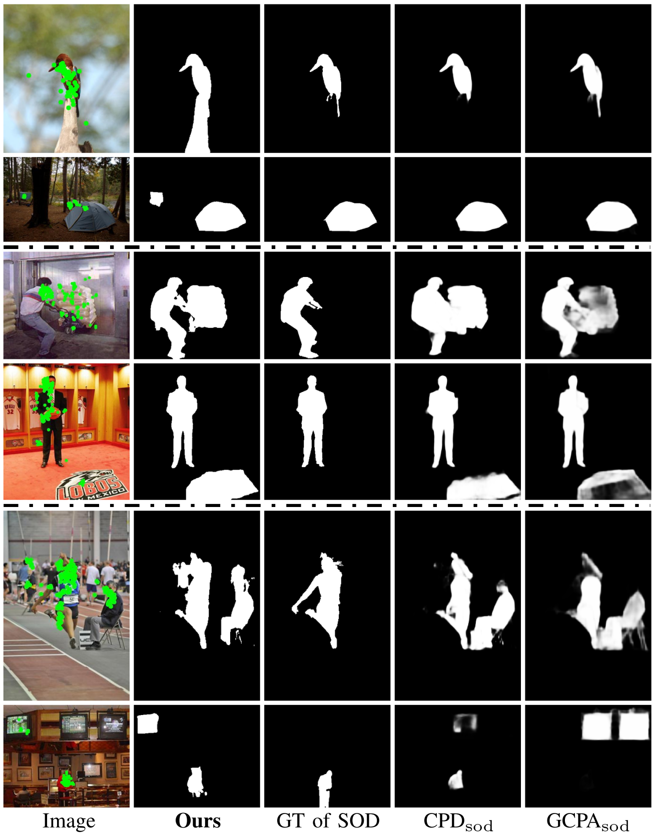


Fig. 9. Visual comparisons between our method, which is proposed for fixation-based object segmentation, and recent state-of-the-art salient object detection methods, including CPD [59] and GCPA [60], on the DUTS-OMRON [46] and PASCAL-S [23] datasets. "GT of SOD" means that the GT is for SOD task. "CPD_{sod}" means the original CPD method for SOD. "GCPA_{sod}" means the original GCPA method for SOD.

cloth that are not found in the GT of SOD and the results of CPD and GCPA. Second, in the 3rd and 4th rows, we find that the results of CPD and GCPA are similar to ours, but different from the GT of SOD. This shows that to some extent, the results of SOD methods CPD and GCPA are consistent with the fixation maps, even if the fixation maps are not exploited in these methods. Third, in the 5th and 6th rows, we can clearly observe that our results are consistent with the fixation points in images, while the other three maps are different. This shows that different SOD methods may cause confusion in some complicated scenes, resulting in inaccurate saliency maps.

Furthermore, we find that the salient objects always appear in the results of our method, while there is ambiguity among different SOD methods, which may highlight different salient objects. So, to improve the accuracy of different SOD methods, we believe that the fixation-based object segmentation can be a pre-processing operation for SOD to determine the salient object proposals.

VII. Conclusion

In this paper, we propose a three-step approach to transform the available fixation prediction dataset OSIE to the PFOS dataset for personal fixations-based object segmentation. The PFOS dataset is meaningful to promote the development of fixations-based object segmentation. Moreover, we present a novel OLBP network with the scheme of bottom-up object localization and top-down boundary preservation to segment the gazed objects. Our OLBP network is equipped with two essential components: the object localization module and the boundary preservation module. OLM is object locator, which is in charge of location analysis of fixations and object enhancement. BPM emphasizes erroneous localization distillation and object completeness preservation. Besides, we provide comprehensive experiments of our OLBP network and other three types of methods on the PFOS dataset, which demonstrate the excellence of our OLBP network and validate the challenges of the PFOS dataset. In our future work, we plan to apply the proposed OLBP network to some eye-control devices, facilitating the lives of patients with hand disability, ALS and polio. In addition, we plan to recruit subjects to collect fixation points and corresponding ground truths on the PASCAL VOC [45] and MS COCO [83] datasets for further exploring personal fixation-based object segmentation.

References

[1] O. Le Meur, A. Coutrot, Z. Liu, P. Rama, A. Le Roch, and A. Helo, "Visual attention saccadic models learn to emulate gaze patterns from childhood to adulthood," *IEEE Trans. Image Process.*, vol. 26, no. 10, pp. 4777–4789, Oct. 2017.

[2] A. Mahdi, M. Su, M. Schlesinger, and J. Qin, "A comparison study of saliency models for fixation prediction on infants and adults," *IEEE Trans. Cognit. Develop. Syst.*, vol. 10, no. 3, pp. 485–498, Sep. 2018.

[3] J. Hewig, R. H. Trippe, H. Hecht, T. Straube, and W. H. R. Miltner, "Gender differences for specific body regions when looking at men and women," *J. Nonverbal Behav.*, vol. 32, no. 2, pp. 67–78, Jun. 2008.

[4] N. Alwall, D. Johansson, and S. Hansen, "The gender difference in gaze-cueing: Associations with empathizing and systemizing," *Personality Individual Differences*, vol. 49, no. 7, pp. 729–732, Nov. 2010.

[5] A. Li and Z. Chen, "Personalized visual saliency: Individuality affects image perception," *IEEE Access*, vol. 6, pp. 16099–16109, 2018.

[6] Y. Xu, S. Gao, J. Wu, N. Li, and J. Yu, "Personalized saliency and its prediction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 12, pp. 2975–2989, Dec. 2019.

[7] Y. Y. Boykov and M.-P. Jolly, "Interactive graph cuts for optimal boundary & region segmentation of objects in N-D images," in *Proc. 8th IEEE Int. Conf. Comput. Vis. ICCV*, Jul. 2001, pp. 105–112.

[8] L. Grady, "Random walks for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 11, pp. 1768–1783, Nov. 2006.

[9] X. Bai and G. Sapiro, "Geodesic matting: A framework for fast interactive image and video segmentation and matting," *Int. J. Comput. Vis.*, vol. 82, no. 2, pp. 113–132, Apr. 2009.

[10] T. Nhat Anh Nguyen, J. Cai, J. Zhang, and J. Zheng, "Robust interactive image segmentation using convex active contours," *IEEE Trans. Image Process.*, vol. 21, no. 8, pp. 3734–3743, Aug. 2012.

[11] T. Vallin Spina, P. A. V. de Miranda, and A. Xavier Falcao, "Hybrid approaches for interactive image segmentation using the live markers paradigm," *IEEE Trans. Image Process.*, vol. 23, no. 12, pp. 5756–5769, Dec. 2014.

[12] N. Xu, B. Price, S. Cohen, J. Yang, and T. Huang, "Deep interactive object selection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 373–381.

[13] J. Liew, Y. Wei, W. Xiong, S.-H. Ong, and J. Feng, "Regional interactive image segmentation networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2746–2754.

[14] Z. Li, Q. Chen, and V. Koltun, "Interactive image segmentation with latent diversity," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 577–585.

[15] S. Mahadevan, P. Voigtlaender, and B. Leibe, "Iteratively trained interactive segmentation," in *Proc. BMVC*, Sep. 2018, pp. 1–12.

[16] W.-D. Jang and C.-S. Kim, "Interactive image segmentation via backpropagating refinement scheme," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5292–5301.

[17] Y. Hu, A. Soltoggio, R. Lock, and S. Carter, "A fully convolutional two-stream fusion network for interactive image segmentation," *Neural Netw.*, vol. 109, pp. 31–42, Jan. 2019.

[18] C. Rother, V. Kolmogorov, and A. Blake, "'GrabCut': Interactive foreground extraction using iterated graph cuts," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 309–314, Aug. 2004.

[19] V. Lempitsky, P. Kohli, C. Rother, and T. Sharp, "Image segmentation with a bounding box prior," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep. 2009, pp. 277–284.

[20] N. Xu, B. Price, S. Cohen, J. Yang, and T. Huang, "Deep grabCut for object selection," in *Proc. Brit. Mach. Vis. Conf.*, 2017, pp. 1–12.

[21] R. Shi, K. N. Ngan, S. Li, and H. Li, "Interactive object segmentation in two phases," *Signal Process., Image Commun.*, vol. 65, pp. 107–114, Jul. 2018.

[22] A. K. Mishra, Y. Aloimonos, L. Fah Cheong, and A. Kassim, "Active visual segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 639–653, Apr. 2012.

[23] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille, "The secrets of salient object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 280–287.

[24] R. Shi, N. K. Ngan, and H. Li, "Gaze-based object segmentation," *IEEE Signal Process. Lett.*, vol. 24, no. 10, pp. 1493–1497, Oct. 2017.

[25] G. Li, Z. Liu, R. Shi, and W. Wei, "Constrained fixation point based segmentation via deep neural network," *Neurocomputing*, vol. 368, pp. 180–187, Nov. 2019.

[26] J. Xu, M. Jiang, S. Wang, M. S. Kankanhalli, and Q. Zhao, "Predicting human gaze beyond pixels," *J. Vis.*, vol. 14, no. 1, pp. 1–20, 2014.

[27] M. Jian and C. Jung, "Interactive image segmentation using adaptive constraint propagation," *IEEE Trans. Image Process.*, vol. 25, no. 3, pp. 1301–1311, Mar. 2016.

[28] T. Wang, J. Yang, Z. Ji, and Q. Sun, "Probabilistic diffusion for interactive image segmentation," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 330–342, Jan. 2019.

[29] M. Sadeghi, G. Tien, G. Hamarneh, and M. S. Atkins, "Hands-free interactive image segmentation using eyegaze," *Proc. SPIE*, vol. 7260, pp. 441–450, Mar. 2009.

[30] G. Kootstra, N. Bergström, and D. Kragic, "Using symmetry to select fixation points for segmentation," in *Proc. IEEE ICPR*, Aug. 2010, pp. 3894–3897.

[31] X. Tian and C. Jung, "Point-cut: Fixation point-based image segmentation using random walk model," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2015, pp. 2125–2129.

[32] N. Khosravan *et al.*, "Gaze2Segment: A pilot study for integrating eye-tracking technology into medical image segmentation," in *Proc. MICCAIW*, Jul. 2017, pp. 94–104.

[33] N. Wang and X. Gong, "Adaptive fusion for RGB-D salient object detection," *IEEE Access*, vol. 7, pp. 55277–55284, 2019.

[34] Y. Wang, X. Zhao, X. Hu, Y. Li, and K. Huang, "Focal boundary guided salient object detection," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2813–2824, Jun. 2019.

[35] Z. Wu, L. Su, and Q. Huang, "Stacked cross refinement network for edge-aware salient object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7263–7272.

[36] J. Zhao, J.-J. Liu, D.-P. Fan, Y. Cao, J. Yang, and M.-M. Cheng, "EGNet: Edge guidance network for salient object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8778–8787.

[37] H. Ding, X. Jiang, A. Q. Liu, N. M. Thalmann, and G. Wang, "Boundary-aware feature propagation for scene segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6818–6828.

[38] W. Wei, Z. Liu, L. Huang, A. Nebout, and O. Le Meur, "Saliency prediction via multi-level features and deep supervision for children with autism spectrum disorder," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Jul. 2019, pp. 621–624.

[39] X. Yang, M.-L. Shyu, H.-Q. Yu, S.-M. Sun, N.-S. Yin, and W. Chen, "Integrating image and textual information in human–robot interactions for children with autism spectrum disorder," *IEEE Trans. Multimedia*, vol. 21, no. 3, pp. 746–759, Mar. 2019.

[40] J. E. Silberg *et al.*, "Free visual exploration of natural movies in schizophrenia," *Eur. Arch. Psychiatry Clin. Neurosci.*, vol. 269, no. 4, pp. 407–418, Jun. 2019.

[41] J. Polec *et al.*, "Detection of schizophrenia spectrum disorders using saliency maps," in *Proc. IEEE 11th Int. Conf. Appl. Inf. Commun. Technol. (AICT)*, Sep. 2017, pp. 1–5.

[42] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep. 2009, pp. 2106–2113.

[43] M. Jiang, S. Huang, J. Duan, and Q. Zhao, "SALICON: Saliency in context," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1072–1080.

[44] K. McGuinness and N. E. O'Connor, "A comparative evaluation of interactive segmentation algorithms," *Pattern Recognit.*, vol. 43, no. 2, pp. 434–444, Feb. 2010.

[45] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Jun. 2010.

[46] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3166–3173.

[47] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, "Salient object detection: A benchmark," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5706–5722, Dec. 2015.

[48] W. Wang, Q. Lai, H. Fu, J. Shen, H. Ling, and R. Yang, "Salient object detection in the deep learning era: An in-depth survey," 2019, *arXiv:1904.09146*. [Online]. Available: http://arxiv.org/abs/1904.09146

[49] Z. Liu, W. Zou, and O. Le Meur, "Saliency tree: A novel saliency detection framework," *IEEE Trans. Image Process.*, vol. 23, no. 5, pp. 1937–1952, May 2014.

[50] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[51] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[52] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *Proc. ICLR*, May 2016, pp. 2–14.

[53] S. Xie and Z. Tu, "Holistically-nested edge detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1395–1403.

[54] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6230–6239.

[55] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[56] H. Zhang *et al.*, "Context encoding for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7151–7160.

[57] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 833–851.

[58] K. Sun *et al.*, "High-resolution representations for labeling pixels and regions," in *Proc. IEEE CVPR*, Jul. 2019, pp. 1–13.

[59] Z. Wu, L. Su, and Q. Huang, "Cascaded partial decoder for fast and accurate salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3902–3911.

[60] Z. Chen, Q. Xu, R. Cong, and Q. Huang, "Global context-aware progressive aggregation network for salient object detection," in *Proc. AAAI*, Feb. 2020, pp. 10599–10606.

[61] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[62] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. AISTATS*, May 2010, pp. 249–256.

[63] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. ACM Int. Conf. Multimedia MM*, 2014, pp. 675–678.

[64] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proc. COMPSTAT*, Aug. 2010, pp. 177–186.

[65] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A new way to evaluate foreground maps," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4548–4557.

[66] R. Margolin, L. Zelnik-Manor, and A. Tal, "How to evaluate foreground maps," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 248–255.

[67] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji, "Enhanced-alignment measure for binary foreground map evaluation," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 698–704.

[68] G. Li, Z. Liu, and H. Ling, "ICNet: Information conversion network for RGB-D based salient object detection," *IEEE Trans. Image Process.*, vol. 29, pp. 4873–4884, Mar. 2020.

[69] G. Li, Z. Liu, L. Ye, Y. Wang, and H. Ling, "Cross-modal weighting network for RGB-D salient object detection," in *Proc. ECCV*, Aug. 2020, pp. 665–681.

[70] N. Otsu, “A threshold selection method from gray-level histograms,” *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-9, no. 1, pp. 62–66, Jan. 1979.

[71] W. Wang and J. Shen, “Deep visual attention prediction,” *IEEE Trans. Image Process.*, vol. 27, no. 5, pp. 2368–2378, May 2018.

[72] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, “Predicting human eye fixations via an LSTM-based saliency attentive model,” *IEEE Trans. Image Process.*, vol. 27, no. 10, pp. 5142–5154, Oct. 2018.

[73] N. Liu, J. Han, T. Liu, and X. Li, “Learning to predict eye fixations via multiresolution convolutional neural networks,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 2, pp. 392–404, Feb. 2018.

[74] Z. Che, A. Borji, G. Zhai, X. Min, G. Guo, and P. Le Callet, “How is gaze influenced by image transformations? Dataset and model,” *IEEE Trans. Image Process.*, vol. 29, pp. 2287–2300, 2020.

[75] K. Fu, C. Gong, I. Y.-H. Gu, and J. Yang, “Normalized cut-based saliency detection by adaptive multi-level region merging,” *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5671–5683, Dec. 2015.

[76] K. Fu, Q. Zhao, and I. Y.-H. Gu, “RefiNet: A deep segmentation assisted refinement network for salient object detection,” *IEEE Trans. Multimedia*, vol. 21, no. 2, pp. 457–469, Feb. 2019.

[77] K. Fu, Q. Zhao, I. Yu-Hua Gu, and J. Yang, “Deepside: A general deep framework for salient object detection,” *Neurocomputing*, vol. 356, pp. 69–82, Sep. 2019.

[78] K. Fu, D.-P. Fan, G.-P. Ji, Q. Zhao, J. Shen, and C. Zhu, “Siamese network for RGB-D salient object detection and beyond,” 2020, *arXiv:2008.12134*. [Online]. Available: http://arxiv.org/abs/2008.12134

[79] K. Fu, D.-P. Fan, G.-P. Ji, and Q. Zhao, “JL-DCF: Joint learning and densely-cooperative fusion framework for RGB-D salient object detection,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3049–3059.

[80] X. Zhou, Z. Liu, C. Gong, and W. Liu, “Improving video saliency detection via localized estimation and spatiotemporal refinement,” *IEEE Trans. Multimedia*, vol. 20, no. 11, pp. 2993–3007, Nov. 2018.

[81] D.-P. Fan, W. Wang, M.-M. Cheng, and J. Shen, “Shifting more attention to video salient object detection,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8546–8556.

[82] W. Wang, J. Shen, J. Xie, M.-M. Cheng, H. Ling, and A. Borji, “Revisiting video saliency prediction in the deep learning era,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 220–237, Jan. 2021.

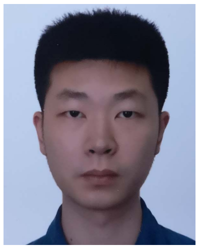
[83] T.-Y. Lin *et al.*, “Microsoft coco: Common objects in context,” in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2014, pp. 740–755.

Gongyang Li (Member, IEEE) received the B.E. degree from Shanghai Normal University, Shanghai, China, in 2016. He is currently pursuing the Ph.D. degree with the School of Communication and Information Engineering, Shanghai University, Shanghai. His research interests include image/video object segmentation and saliency detection.

Zhi Liu (Senior Member, IEEE) received the B.E. and M.E. degrees from Tianjin University, Tianjin, China, in 1999 and 2002, respectively, and the Ph.D. degree from the Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai, China, in 2005. From August 2012 to August 2014, he was a Visiting Researcher with the SIROCCO Team, IRISA/INRIA-Rennes, France, with the support by EU FP7 Marie Curie Actions. He is currently a Professor with the School of Communication and Information Engineering, Shanghai University, Shanghai. He has published more than 200 refereed technical articles in international journals and conferences. His research interests include image/video processing, machine learning, computer vision, and multimedia communication. He was a TPC member/Session Chair of ICIP 2017, PCM 2016, VCIP 2016, ICME 2014, and WIAMIS 2013. He co-organized special sessions on visual attention, saliency models, and applications at WIAMIS 2013 and ICME 2014. He is also an Area Editor of *Signal Processing: Image Communication*. He has served as a Guest Editor for the Special Issue on Recent Advances in Saliency Models, Applications and Evaluations in *Signal Processing: Image Communication*.

Ran Shi received the B.S. degree in electronic science and technology from the Changshu Institute of Technology in 2009, the M.S. degree in signal and information processing from Shanghai University in 2012, and the Ph.D. degree in electronic engineering from The Chinese University of Hong Kong (CUHK) in 2017. He joined CUHK as a Research Assistant in 2012. He is currently an Assistant Professor with the School of Computer Science and Engineering, Nanjing University of Science and Technology. His research interests include object segmentation, visual quality evaluation, interactive image segmentation, and salient object detection.

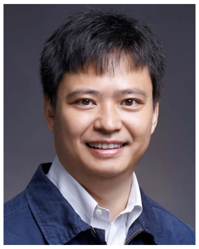
Zheng Hu received the B.E. degree from Shanghai University, Shanghai, China, in 2018, where he is currently pursuing the M.E. degree with the School of Communication and Information Engineering. His research interests include instance segmentation and interactive image segmentation.



Weijie Wei received the B.E. degree from Shanghai University, Shanghai, China, in 2018, where he is currently pursuing the M.E. degree with the School of Communication and Information Engineering. His research interests include deep learning and saliency prediction.

Yong Wu received the B.E. degree from the Anhui Science and Technology University, Bengbu, China, in 2015, and the M.S. degree from Shantou University, Shantou, China, in 2018. He is currently pursuing the Ph.D. degree with School of Communication and Information Engineering, Shanghai University, Shanghai, China. His research interests include machine learning, visual tracking, and gaze estimation.

Mengke Huang received the B.E. degree from Shanghai Normal University, Shanghai, China, in 2014. He is currently pursuing the Ph.D. degree with the School of Communication and Information Engineering, Shanghai University, Shanghai. His research interests include saliency detection and deep learning.



Haibin Ling received the B.S. and M.S. degrees from Peking University in 1997 and 2000, respectively, and the Ph.D. degree from the University of Maryland, College Park, in 2006. From 2000 to 2001, he was an Assistant Researcher with Microsoft Research Asia. From 2006 to 2007, he worked as a Post-Doctoral Scientist with the University of California Los Angeles. In 2007, he joined Siemens Corporate Research as a Research Scientist; then, from 2008 to 2019, he worked as a Faculty Member of the Department of Computer Sciences, Temple University. In fall 2019, he joined Stony Brook University as a SUNY Empire Innovation Professor at the Department of Computer Science. His research interests include computer vision, augmented reality, medical image analysis, and human–computer interaction. He received the Best Student Paper Award from ACM UIST in 2003, the NSF CAREER Award in 2014, the Yahoo Faculty Research and Engagement Program Award in 2019, and the Amazon AWS Machine Learning Research Award in 2019. He has served as an Area Chair various times for CVPR and ECCV. He also serves as an Associate Editor for several journals, including the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE (PAMI), *Pattern Recognition* (PR), and *Computer Vision and Image Understanding* (CVIU).