Neurocomputing 445 (2021) 255–266

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Personalized image observation behavior learning in fixation based personalized salient object segmentation

Ran Shi [a], Gongyang Li [b], Weijie Wei [b], Xiaofei Zhou [c,*], Zhi Liu [b]

[a] school of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China
[b] school of Communication and Information Engineering, Shanghai University, Shanghai, China
[c] Institute of Information and Control, School of Automation, Hangzhou Dianzi University, Hangzhou, China

ARTICLE INFO

Article history:
Received 3 October 2020
Revised 8 March 2021
Accepted 11 March 2021
Available online 18 March 2021
Communicated by Zidong Wang

Keywords:
Fixation
Salient object segmentation
Meta-learning

ABSTRACT

Fixation as representation of one viewer's attention are very intuitive to reflect the viewer's observation procedure. The viewer's observation behavior can be further revealed by analyzing fixations features. In this paper, we propose a fixation based personalized salient object segmentation method involving personal observation behavior learning. Concretely, we design three neural networks and deploy a meta-learning method. The first network is a base segmentation network that can be converted into a meta-segmentation network by meta-learning. The meta- segmentation network can learn one viewer's observation behavior from only one sample and then generates the viewer's segmentation network to segment the other samples. Moreover, a fusion network plays an important role in alleviating an unsuitable transmission problem and generating a final segmentation result. The experimental results demonstrate the reasonability of our observation behavior learning and the effectiveness of the three proposed neural networks.

1. Introduction

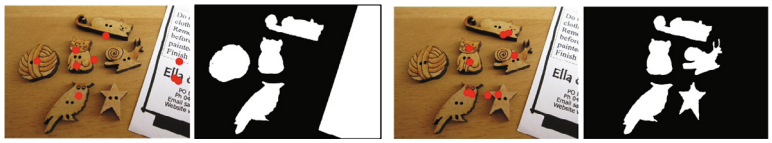
Salient object segmentation aims to extract the most eye-attracting object regions in a given scene [1–3]. It can assist subsequent image processing applications to fulfill object-aware analysis and manipulation, such as object-aware image cropping [4] and compression [5]. In fact, heterogeneity in different viewers' interest preferences for one image has been widely recognized in psychology [6]. Ref. [7] also verified that different viewers have different salient objects in certain scenes with multiple objects. Therefore, personalized salient object segmentation is essential to make object-aware image applications more individualized by designing tailored processing to cater to different viewers' needs.

With the development of eye tracking technique, an eye tracker device can record one viewer's fixation information on certain images in real time [8]. Since fixation is a kind of representation of attention, personalized fixation can be treated as an intuitive and convenient input to drive personalized salient object segmentation as shown in Fig. 1. This implicit input mode is different from the mode adopted in traditional interactive object segmentation, which draws some scribbles on the object and background respectively [9]. Compared with the labeled scribbles adopted in the traditional interactive object segmentation, fixations are unlabeled and cannot directly distinguish their belonging regions. Fixations actually reflect one viewer's whole observation procedure on the image so that background and non-salient object regions can also obtain fixations. Therefore, this indicative ambiguity is one problem for fixation based personalized salient object segmentation.

Additionally, even if different viewers are attracted by the same salient object, another problem induced by fixations is that different viewers still have their own observation behaviors. Examples of two images with two viewers' fixations are shown in Fig. 2. The fixation distributions of these two viewers in the same image are quite different. One viewer focuses on high semantic regions only due to concentrated fixations. Conversely, for the other viewer, the observation extent is larger, as indicated by the scattered fixations. Therefore, it is difficult to design a universal model for all viewers to segment their salient objects. However, every coin has two sides. We can also see that since these two images are similar to some extent, i.e., one prominent object, these two viewers adopt similar observation behaviors on these similar scenes. Therefore, if one viewer's personalized observation behavior on a certain image can be learned, fixation based segmentation methods can better understand and utilize this viewer's fixations. This means that the learned observation behavior may alleviate

* Corresponding author.
E-mail address: zxforchid@hdu.edu.cn (X. Zhou).

https://doi.org/10.1016/j.neucom.2021.03.042
0925-2312/



(a) (b)

Fig. 1. Examples of different personalized fixation distributions on one same image and corresponding salient object segmentation masks. Fixations are labelled as red dots. (a) and (b) represent two different viewers.

the indicative ambiguity problem and improve this viewer's salient object segmentation in other similar scenes.

For known viewers, we can collect their lots of samples and then train their models [7]. However, it is invalid in practice when viewers are unknown in advance. How to quickly learn an unknown viewer's observation behavior is challenging due to limited samples. Meta-learning approaches [10–12] provide a possible paradigm for solving the problem of learning from a few samples. Meta-learning, also known as learning-to-learn, was first studied in [13] and has regained attention after entering the deep learning era. It aims to learn on a bunch of similar tasks to maximize its adaptation performance to all tasks. The key to meta-learning is to train a meta-learner as a learned optimizer, which can offer flexible learning rules to adapt quickly from a few task-specific training samples and apply the rules to new and similar test samples. For example, [10] trained a meta-learner long short-term memory (LSTM) component [14] to learn the update rule for training a model; Ref. [11,12] focused on learning model initialization and proposed fairly general optimization algorithms that were compatible with any model that learns through gradient descent. Therefore, we seek to take advantage of meta-learning to learn an unknown viewer's observation behavior by only one sample.

In this paper, we involve personalized image observation behavior learning by integrating meta-learning into fixation based personalized salient object segmentation. Compared with previous related works, the main contributions of our work are summarized as follows:

1. Personalized image observation behavior learning is involved in our segmentation method so that one viewer's similar observation behaviors in similar scenes can be transmitted to assist this viewer's salient object segmentation.
2. We interpret features of the observation behavior as a compromise of features of the fixation distribution and features of the fixation depending on the image.
3. We design three branches and take advantage of the property of the convolutional long short-term memory component to extract the observation behavior features and fuse them with features of the image.
4. To handle a possible unsuitable transmission of the observation behavior, we consider weighted samples and a measure of the reliability of the meta adaptation in the meta-training stage and additionally design a fusion network.

This paper is organized as follows. Section 2 describes our personalized image observation behavior learning in details. The implementation of fixation based personalized salient object segmentation is introduced in Section 3. Experimental results are presented in Section 4. We conclude our paper in Section 5.

2. Related work

From the perspective of underlying mechanisms to obtain attention, attention models can be classified into two types: one is driven by target-object goals, and the other is inspired by stimuli of visual input [15–18]. Thus, the fixation as a kind of representation of the attention also has different characteristics accordingly, which are applied in different tasks. For the former, the distribution of fixations is highly correlated with the target-object goals [18]. Many works [18–24] attempt to predict target-object goal directed fixations to assist the object visual searching task. [19] bridged fixations with the searching tasks of specific object categories, i.e., a microwave and a clock. [20] predicted fixation' directions of several individuals within social scenes to infer their one common target-object goal. For more general object categories, [21,22] used computational models that treated target spatial relationships [21] and global scene context [22] as important factors affecting visual searching. Some deep network models simulated eye movement theories or learning techniques to predict searching fixations, such as biologically plausible mechanisms of inhibition

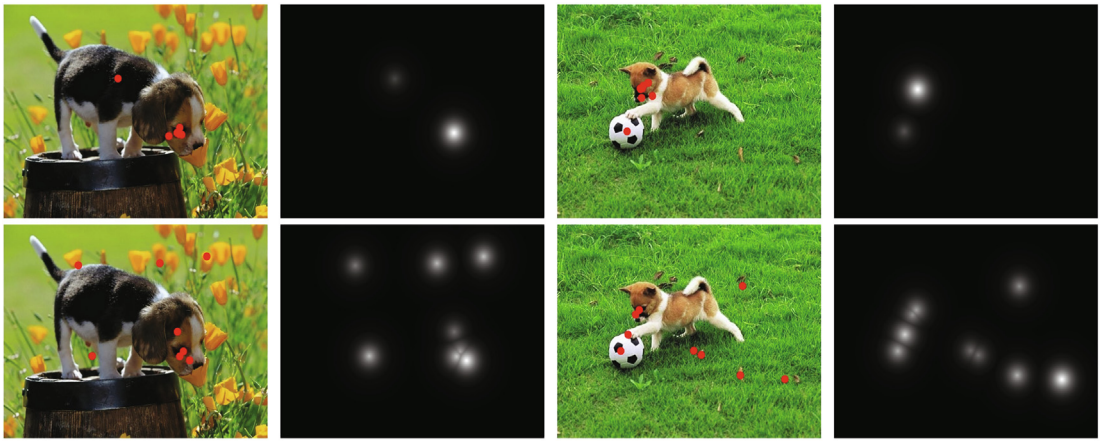


Fig. 2. Examples of two different viewers' fixation distributions. Each row shows one viewer's fixations indicated as red dots in two images and corresponding fixation maps.

of return [23] and biased-competition theory [24]. In [18], the viewer's dynamic contextual belief maps of object locations were learned by inverse reinforcement learning and then used to predict behavioral scan paths for multiple categories. In [25,26], fixations driven by target-object goals were also used to guide interactive segmentation. [25] collected fixations of radiologists when they diagnosed lesions in lung CT images. These fixations are treated as seeds of lesion. Then, background seeds can be further found by using gradient information of the CT image. Thus, lesions can be segmented by any seed-based interactive segmentation method. [26] developed a new graphical interactive segmentation interface controlled by fixations. It requests users to look at their target objects and the background respectively so that these collected fixations can be automatically labeled as object seeds and background seeds.

If there are no target-object goals for viewers, free-viewing fixation is inspired by stimuli that is visual information of interest. In images and videos, the visual information of interest is generally represented as the salient object that can draw more attention from viewers compared with other regions. Many works focus on either the free-viewing fixation or the salient object [27–30]. For the fixation, [28] captured hierarchical saliency information by a skip-layer based network structure to predict human eye fixation with view-free scenes. In [29], a DHF1K benchmark predicting fixations during dynamic scene free-viewing was reported and an attentive CNN-LSTM architecture was proposed that leveraged both static and dynamic fixation data to encode static attention into dynamic saliency representation learning. For the salient object, [27] presented a salient object detection method that integrated both top-down and bottom-up saliency inference iteratively and cooperatively. Ref. [30] provided a comprehensive survey of deep learning based salient object detection methods in terms of network architecture, level of supervision, learning paradigm, object-/instance-level detection and performance under different datasets and attribute settings. In this paper, we focus on the relationship between fixations and the salient object. In [31,32], it leveraged fixation prediction for detecting salient objects by an attentive saliency network. The fixation map was learned from the upper layers of the attentive saliency network and further utilized for teaching the network to detect the salient object. In [17], a biologically-inspired dynamic visual attention prediction module was developed and used to guide its attention-guided object segmentation module for fine-grained unsupervised video object segmentation. Actually, the salient object focused by these works in different scenes are universal salient objects, which are common salient objects of major viewers. This means that the fixations/attention predicted by these methods are also universal. However, [7] claimed that heterogeneity in the saliency preference of different viewers has been widely recognized in psychology [6]. This verified that individuals exhibit heterogeneous fixation patterns when viewing an identical scene containing multiple salient objects and presented a multi-task convolutional neural network framework for predicting the discrepancy between personalized saliency maps and one universal saliency map. Inspired by [7], our work can be treated as a personalized, fixation-driven salient object segmentation method. One intuitive method is to collect one viewer's own fixations by the eye tracker and directly use them as input to direct the segmentation, similar to [25,26]. However, free-viewing fixations induce indicative ambiguity because they are unlabeled and cannot be used to directly distinguish their belonging regions. Some methods attempt to infer object and background information from the fixations [4,33,34] in some simple scenes. In [4], one image was segmented into several superpixels which can be further divided into "object seed", "background seed" and "unknown region" depending on the distribution of the fixations on them. In [34], the centroid of fixations and the mean distance of all fixations from this centroid were used to choose the segmentation seeds. [33] took advantage of saliency maps as auxiliary information. Following the above ideas, once the object and background cues can be estimated, object and background models can be constructed. Alternatively, Ref. [3,35] proposed selection-aware methods that utilize object proposals to generate several candidate segmentation results. Then, they extracted some hand-crafted features of the fixation distributions on each candidate result and estimated a score to indicate the possibility that it belongs to salient objects. Thus, the score ranking can be used to "select" the final segmentation result. In addition, Li et al. [36] concatenated the image and personalized fixation map as the input and proposed a convolutional neural network based model to simulate the human visual system to segment the objects. However, it is still difficult to design a uniform model for all viewers to segment their salient objects. None of the above mentioned methods consider the role played by one viewer's own observation behavior in the personalized salient object segmentation.

3. Personalized image observation behavior learning

Our proposed personalized observation behavior learning depends on a fixation-driving segmentation network and a meta-learning method. Therefore, we first introduce the architecture of our segmentation network and the meta-learning method in this section.

3.1. Input representation

Since our network services to fixation based salient object segmentation, the image and the fixation as the input of our neural network are necessary. Although the eye tracker records fixations' durations and their positions as points, it does not mean that the viewer merely looks at these points in an image. It should be a certain attention extent induced by one fixation. Therefore, we convert the fixations' information into an fixation map FM to estimate an attention degree of each pixel. For one pixel x in the FM, its attention degree $FM(x)$ is:

$$FM(x) = \max_{y \in \Omega_{fix}} \left(\frac{T(y)/T_{max}}{1 + exp(D(x,y)/\sigma)} \right) \quad (1)$$

where y is one fixation in the fixation set Ω_{fix}. According to the distance $D(x,y)$ from x to y and the duration $T(y)$ of y, if $D(x,y)$ is shorter and $T(y)$ is longer, it means that x also draws much attention. T_{max} indicates the longest duration of all fixations and σ controls the attention extent stimulated by the exponential function. σ is set to 20 in our experiments. $FM(x)$ is the maximum value among all pairs of x and y. After normalization and scaling the attention degree into $[0, 255]$, the FM can be generated. Meanwhile, we also introduce a semantic map (SM) as another input because semantic information is an important factor to analyze images' scenes and influence the viewer's fixations. We derive the feature maps "conv5_3" from the pre-trained VGG-16 network [37], which involve high level semantic information as disclosed in [38]. Since there are 512 channels of "conv5_3", we select the maximum value of each channel and normalize them to compose our SM. Thus, it can indicate rough location distribution of all high semantic regions in an image. One example of an original image and its semantic map is illustrated in Fig. 3.

3.2. Neural network architecture

There are two parts in our neural network as illustrated in Fig. 4. The former part extracts image features and the viewer's observation behavior, and then fuses them. The latter part is designed for

(a) (b)

Fig. 3. One example of an original image and its semantic map. (a) Original image, (b) Semantic map.

fulfilling the segmentation task. For the former part, since the viewer's observation behavior is dependent on the image, as discussed before, it cannot be merely learned from the fixation map only. Features extracted from the fixation map alone merely describe characteristics of the distribution of the fixations. This means that they cannot well reflect the relationship between fixations and the corresponding image. Therefore, the fixation features should not be extracted without the accompaniment of the image. However, if the viewer's observation behavior is learned from a combination of the image and the fixation map, extracted features may excessively focus on the relationship between fixations and image specific details. This results in limitations on the generality of observation behavior learning. In our opinion, the observation behavior features that we pursue should be a compromise of fixation distribution features and image dependent features. Therefore, we design two branches for the observation behavior feature extraction: one is a "high" branch whose input X_h is the fixation map *FM* only and the other is a "low" branch whose input X_l is a five-channel tensor by concatenating the image *I*, the fixation map *FM* and the semantic map *SM*. Meanwhile, we also add another "image" branch whose input X_{img} is a four-channel tensor with image *I* and semantic map *SM* to extract features of the image. For each branch, we design the same structure which is composed of several standard convolutional layers and dilated convolutional layers as shown in Fig. 4. The dilated convolutional is used to enlarge the network's receptive filed and maintain the input resolution [39]. These three branches can be formulated as:

$$Y_h = CNN_h(X_h) \tag{2}$$
$$Y_l = CNN_l(X_l) \tag{3}$$
$$Y_{img} = CNN_{img}(X_{img}) \tag{4}$$

where CNN_h, CNN_l and CNN_{image} are the convolutional layers of high branch, low branch and image branch. *Y* indicates the corresponding output of each branch.

Then, we adopt a convolutional long short-term memory (ConvLSTM) component [40] to extract the feature of observation behavior and further fuse them with image features. The fused features as output is fed to the latter segmentation part. Similar to traditional gated LSTM [14], the ConvLSTM uses the memory cells including the Cell state (C state) and the Hidden state (H state), and four gates $\mathbf{i}$, $\mathbf{f}$, $\mathbf{c}$, $\mathbf{o}$ to control information flow. It extends traditional fully connected LSTM by substituting dot products with convolutional operations in the LSTM equations, which can preserve the spatial information of features. Besides the input of the ConvLSTM, the Cell state and the Hidden state can also be treated as two hidden inputs. So, these three inputs can exactly correspond to the outputs of our three feature extraction branches. Especially, according to the inside structure of the ConvLSTM shown in Fig. 4, the hidden state is firstly concatenated with the input to further extract features. Then, they are fused with the cell state to generate the output. Therefore, features extracted by our low branch and high branch are fed to the input and the hidden state of the ConvLSTM respectively. Thus, the observation behavior features can be further extracted and fused with image features provided by the cell state. The whole procedure mentioned above can be formulated as below:

$$\mathbf{i} = \sigma(W_{\mathbf{i}}^{Y_{img}} * Y_{img} + W_{\mathbf{i}}^{Y_h} * Y_h + b_{\mathbf{i}}) \tag{5}$$
$$\mathbf{f} = \sigma(W_{\mathbf{f}}^{Y_{img}} * Y_{img} + W_{\mathbf{f}}^{Y_h} * Y_h + b_{\mathbf{f}}) \tag{6}$$
$$\mathbf{o} = \sigma(W_{\mathbf{o}}^{Y_{img}} * Y_{img} + W_{\mathbf{o}}^{Y_h} * Y_h + b_{\mathbf{o}}) \tag{7}$$
$$C_o = \mathbf{f} \circ Y_l + \mathbf{i} \circ tanh(W_{\mathbf{c}}^{Y_{img}} * Y_{img} + W_{\mathbf{c}}^{Y_h} * Y_h + b_{\mathbf{o}}) \tag{8}$$
$$H_o = \mathbf{o} \circ tanh(C_o) \tag{9}$$

where "$*$" denotes the convolution operator and "$\circ$" represents Hadamard product. σ and *tanh* are the activation functions of logistic sigmoid and hyperbolic tangent. *W*s and *b*s are the learned weights and biases. H_o is also the output of the ConvLSTM.

For the segmentation part, it utilizes the output H_o from ConvLSTM to fulfill the fixation based salient object segmentation task. We simply use three standard convolution layers CNN_{seg} where the last layer adopts the sigmoid activation function to map the segmentation result *R* into $[0, 1]$. The segmentation part can be formulated as:

$$R = CNN_{seg}(H_o) \tag{10}$$

The details of the parameters of each layer in our segmentation network is also listed in the Fig. 4.

3.3. Meta-learning

As the network architecture described above, our segmentation network can be used as a "base segmentation network" (BSN) by

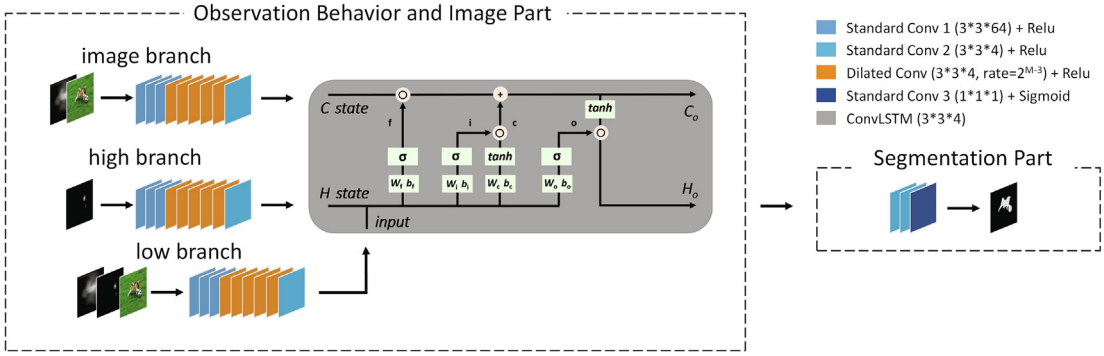


Fig. 4. The architecture of our fixation based personalized salient object segmentation network. *M* is the NO. of the layer.

training accordingly. In order to convert this BSN into meta-segmentation network (MSN) which can quickly learn one viewer's observation behavior and then generates this viewer's segmentation network, MSN is initialized by BSN and trained by Model-Agnostic Meta-Learning (MAML) [11]. In the meta-training process, one task $\mathscr{T}$ of one viewer is denoted as $\{S_m, S_1, \ldots S_K\}$, where S represents a sample. This viewer's own segmentation network is derived from MSN by learning his own observation behavior from S_m. This process is also called meta-adaptation. Then, the segmentation network is utilized to handle the viewer's other K samples in the meta-test process and the average loss is accumulated. The accumulated loss of several viewers' tasks is used to update MSN for enhancing its generalization. Since MSN focuses on learning the viewer's observation behavior, we merely update the parameters of the high branch, the low branch and the ConvLSTM which are related to the extraction of the observation behavior features and their fusion with image features. It means that other parameters in MSN are frozen during the whole meta-training.

4. Implementation of the fixation based personalized salient object segmentation

By the personalized observation behavior learning, our segmentation method attempts to utilize similar observation behaviors in similar scenes to detour the fixation's indicative ambiguity problem and enhance the segmentation quality. However, in real practice, our method should overcome an unsuitable transmission problem in the meta-training and the final segmentation result generation.

4.1. Unsuitable transmission problem

In real practice, if scenes of images are not similar, one potential problem is that one viewer's observation behavior learned from his one sample may not always be suitably transmitted to other samples due to rich diversity of image contents. For example, if there are multiple toys in an image, the learned observation behavior is not suitably transmitted to another image containing only one deer as shown in Fig. 5. The unsuitable transmission may severely degrade the segmentation result. In order to alleviate this problem, we additionally design another fusion network. This network is used to fuse the segmentation results of the base segmentation network and the meta-segmentation network for inferring a good final result. We concatenate the original image with the segmentation results of the base segmentation network and the meta-segmentation network as the input of the fusion network. The structure of the fusion network is similar to that of each branch in our segmentation network. The slight difference is that we add three standard convolutional layers (whose kernel sizes are all $3 \times 3 \times 64$) followed six dilated convolutional layers. Its final result is also generated by the sigmoid activation function.

(a) (b)

Fig. 5. One example of unsuitable transmission problem. (a) Fixations on the multiple objects, (b) Fixations on the single object.

4.2. Training

Our whole training procedure can be divided into three stages: the base segmentation network training, the meta-segmentation network training, and a fusion network training. In the first stage, we train our neural network as the base segmentation network without distinguishing viewers and their samples. This segmentation network can generate segmentation results as references and initialize the meta-segmentation network.

In the second stage, the base segmentation network is converted into the meta-segmentation network by the meta-training. Thus, our meta-segmentation network can generate one viewer's own segmentation network by learning the viewer's observation behavior. In the meta-segmentation network training, as mentioned above, one viewer's one task $\mathscr{T}$ includes one S_m and other K samples. We compose one viewer's various tasks by randomly combing the viewer's different samples. All tasks of n viewers in the training consist of a task distribution $p(\mathscr{T})$. The training process of one epoch is illustrated in Table 1. $\boldsymbol{\theta}_m$ is the parameters of the meta-segmentation network which is initialized by those of the base segmentation network $\boldsymbol{\theta}_b$. In one epoch, we first sample one batch of n tasks from $p(\mathscr{T})$ where n corresponds to those n viewers. For ith viewer's task, we utilize the viewer's S_m to generate the viewer's segmentation network whose parameters $\boldsymbol{\theta}_i$ are initialized by $\boldsymbol{\theta}_m$ and then are updated η steps. $\mathscr{L}$ represents the loss function i.e. the dice coefficient. Then the viewer's segmentation network is to segment other K samples, the average loss are calculated and corresponding gradients are accumulated. Finally, $\boldsymbol{\theta}_m$ is updated by the accumulative gradients of the batch. Notice that the updating of $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}_m$ are only valid on the parameters of the fixations related branches and the ConvLSTM as discussed before. Moreover, due to the unsuitable transmission problem, when $\mathscr{L}(S_k, \boldsymbol{\theta}_i)$ is large, we cannot tell the reason is that our meta-segmentation network is not trained well or the transmission itself is not unsuitable. Therefore, if $\mathscr{L}(S_k, \boldsymbol{\theta}_i) < \mathscr{L}(S_k, \boldsymbol{\theta}_b)$, it means that this transmission indeed improve the segmentation quality compared against the reference result, we introduce a coefficient c to increase this sample's impact on the gradient accumulation. c correlates to a relative loss gain between $\mathscr{L}(S_k, \boldsymbol{\theta}_i)$ and $\mathscr{L}(S_k, \boldsymbol{\theta}_b)$. So, when we gradually strengthen our meta-segmentation network's ability, the updating of our meta-segmentation network more depends on those transmittable samples. Meanwhile, we also consider the quality of the meta-adaptation step. If S_m itself cannot be learned well in the meta-adaptation, the subsequent transmission and the induced loss make no sense. Therefore, we treat $1 - \mathscr{L}(S_m, \boldsymbol{\theta}_i)$ as a weight to measure the reliability of the meta-adaptation using S_m.

The fusion network is trained in the third stage. We perform the meta-training stage again to obtain the segmentation result by the trained meta-segmentation network without updating. Each sample is also segmented by the base segmentation network to get a referable result. Then, we concatenate the original image with these two segmentation results as one training sample and feed to our fusion network for training.

5. Experiments

5.1. Dataset

To the best of our knowledge, there is no public dataset specifically designed for the fixation based personalized salient object segmentation. Therefore, we re-forge a related public dataset OSIE [41] to produce an OSIE-Personalized Salient Object Segmentation dataset (OSIE-P). OSIE dataset provides 15 viewers' fixation information of 700 images recorded by the eye tracker device (Eyelink

Table 1
Meta-training for the fixation based personalized salient object segmentation in one epoch.

Input: task distribution $p(\mathscr{T})$, learning rate $\alpha = 10^{-4}$, $\beta = 10^{-4}$

Output: $\boldsymbol{\theta}_m$
while not done **do**
  Sample one batch of tasks $\{\mathscr{T}_i\}_{i=1}^{n}$ from $p(\mathscr{T})$;
  $\nabla_{\boldsymbol{\theta}_m} \leftarrow \mathbf{0}$
  for $i \leftarrow 1, n$ **do**
    $\boldsymbol{\theta}_i \leftarrow \boldsymbol{\theta}_m$
    $\nabla_{\boldsymbol{\theta}_i} \leftarrow \mathbf{0}$
    for $j \leftarrow 1, \eta$ **do**
      $\boldsymbol{\theta}_i \leftarrow \boldsymbol{\theta}_m - \alpha \cdot \nabla \mathscr{L}(S_m, \boldsymbol{\theta}_i)$
    end for
    for $k \leftarrow 1, K$ **do**
      if $\mathscr{L}(S_k, \boldsymbol{\theta}_i) < \mathscr{L}(S_k, \boldsymbol{\theta}_b)$
        $c \leftarrow \frac{2 * \mathscr{L}(S_k, \boldsymbol{\theta}_b)}{\mathscr{L}(S_k, \boldsymbol{\theta}_b) + \mathscr{L}(S_k, \boldsymbol{\theta}_i)}$
      else
        $c \leftarrow 1$
      end if
      $\nabla_{\boldsymbol{\theta}_i} \leftarrow \nabla_{\boldsymbol{\theta}_i} + c \cdot \nabla \mathscr{L}(S_k, \boldsymbol{\theta}_i)$
    end for
    $\nabla_{\boldsymbol{\theta}_m} \leftarrow \nabla_{\boldsymbol{\theta}_m} + (1 - \mathscr{L}(S_m, \boldsymbol{\theta}_i)) \frac{\nabla_{\boldsymbol{\theta}_i}}{K}$
  end for
  $\boldsymbol{\theta}_m \leftarrow \boldsymbol{\theta}_m - \beta \cdot \nabla_{\boldsymbol{\theta}_m}$
end while

1000) and manual labeled masks of all objects in these images. In the OSIE-P, if one object in an image v can obtain more than Th_{uv} fixations of one viewer u, this object is chosen as this viewer's salient object. The threshold Th_{uv} is calculate as $[N_{uv} \setminus N_R]$, where N_{uv} is the number of one user u's all fixations in the image v. If N_{uv} is less than a reference number N_R, the threshold is set to 1. By analyzing the numbers and the distributions of the fixations of all users in the whole dataset, we deliberately set N_R to 10, which is a proper trade-off between the numbers of one user's fixations and salient objects. After the thresholding, we can generate a viewer u's ground truth of salient object mask map for image v. Correspondingly, there are 15 different ground truths for one image according to different viewers' fixations. A tuple of one image, one viewer's fixation map and corresponding ground truth is treated as one sample.

In order to verify the heterogeneity of different viewers in terms of their salient objects in the OSIE-P, each viewer's ground truth is compared with other 14 viewers' ground truths one by one for one same image. Therefore, for one viewer and 700 images, there are 9800 comparison pairs. We count the number of comparison pairs whose salient objects are not identical. If two viewers have same salient objects, we also assess the similarity of their fixations' distributions by calculating the dice coefficient of their fixation maps. A larger dice coefficient indicates higher similarity. The evaluation results of all 15 viewers are shown in Fig. 6. We can see that each viewer has about 70% comparison pairs including different salient objects. For the rest comparison pairs, the average dice coefficient is round 0.39 which indicates relatively low similarity of their fixation maps. On the one hand, for most comparison pairs including the same salient object, the reason is that there is one dominant object in the image. On the other hand, for similar fixation maps of different viewers, their fixations mostly concentrate on certain special semantic regions (e.g. face) which has been concluded in [7] but fixations' distributions on other regions are various. The lowest percentage and the highest average dice coefficient is 0.683 and 0.405 respectively. The evaluation results demonstrate the heterogeneity of different viewers in terms of their salient objects and fixations' distributions in the OSIE-P.

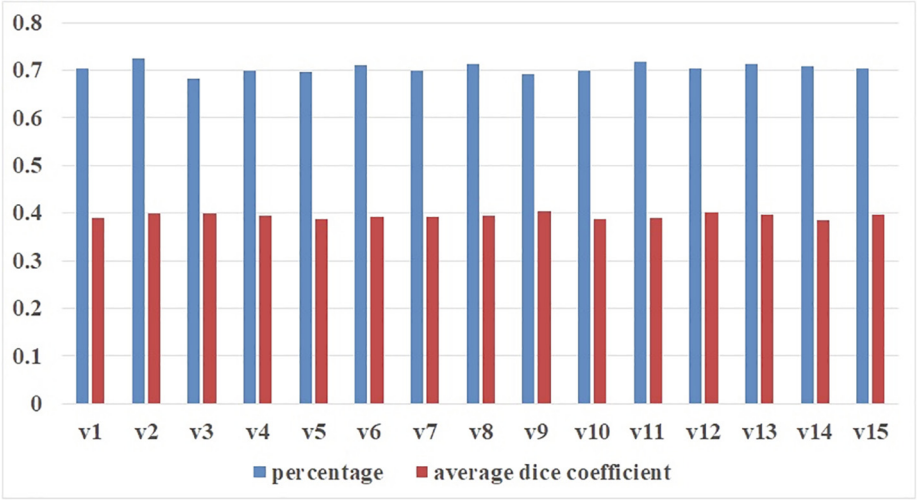


Fig. 6. The percentage of comparison pairs with different ground truths indicated by the blue bar and the average dice coefficient of rest comparison pairs' fixation maps given by the red bar of each viewer "v" in the OSIE-P.

5.2. Implementation details

In the OSIE-P, we randomly select 550 images of 10 viewers with 5500 samples as a training set, 40 images of 2 viewers with 80 samples as a cross validation set and 110 images of 3 viewers with 330 samples as a test set. For the base segmentation network training stage, the original image, the feature map and the semantic map are resized to 200×150. We use the dice coefficient as the loss function. Our base segmentation network is trained using Adam [42], with single sample and learning rate 0.0001. The training proceeds for 20 epochs. In the meta-segmentation network training, K which represents the number of samples in one task is set to 20. n is 10 to correspond to those 10 viewers in the training set. The updating time η in the meta-adaptation is set to 10. The training process iterates for 20 epochs. For the fusion network training, it is also trained using the dice coefficient and Adam, and proceeds for 40 epochs. The final binary segmentation result is generated by a threshold as 0.5.

5.3. Overall performance

In order to evaluate our proposed method's performance, we tested it on the OSIE-P test set. We randomly selected 10 original images with each viewer's fixation maps in the test set as candidate meta-adaptation samples. Using one viewer's one meta adaptation sample, the remaining 100 samples can be accordingly segmented for the test. To evaluate the reliability of our method, we performed this procedure 10 times corresponding to 10 candidate meta -adaptation samples for a total of three viewers. Thus, the average Jaccard Index (i.e., IoU) of all 3,000 segmentation results was used to indicate our method's overall accuracy. In addition, to further analyze the contributions of our three networks to the overall performance, we evaluated the accuracies of the base segmentation network only (BSN) and the meta-segmentation network (MSN) without the fusion network (FN). For MSN, we also assessed its performance when the regular MAML was used in the meta-training step without the weighted samples and the measure of the reliability of the meta-adaptation.

As shown in Table 2, the accuracy of the base segmentation network can be treated as a baseline. Without the fusion network, all samples are segmented by the meta-segmentation network after corresponding meta-adaptations. Its worse performance demonstrates that observation behavior transmission cannot always be done for all samples. Compared with the regular MAML, which enforces the network to fit the unsuitable transmission samples in the meta-training stage, our strategy of the weighted samples and the measure of the reliability of the meta -adaptation can screen the unsuitable transmission samples and decrease their negative influence in the meta-training stage to some extent. Therefore, the performance of our modified MAML is better than that of the regular MAML. However, unsuitable transmission can still distort segmentation quality. Therefore, the fusion network is important to take advantage of the results from BSN and MSN to compensate for possible unsuitable transmissions. Fig. 7 shows some segmentation results of our method. In these examples, similar observation behaviors in the meta -adaptation sample and the test sample are represented in diverse scenes, including one single person, two separate objects and multiple objects. We can see that suitable transmission of the observation behaviors can make our meta segmentation network generate better segmentation results. However, once an unsuitable transmission occurs due to the different observation behaviors of objects of various sizes as illustrated in the last row, the results generated by MSN can be considerably degraded. Therefore, the fusion network plays a role in exploring the final result as well as possible.

Table 2
The overall performance of our method and the ablation study.

Method	*IoU* ↑
BSN	0.584
MSN(regular MAML) w/o FN	0.421
MSN w/o FN	0.543
Our (BSN + MSN + FN)	0.622

Moreover, we further enumerate the average accuracies of each viewer's 100 test samples under each meta-adaptation sample as shown in Fig. 8. The numbers along the horizontal axis represent those 10 candidate meta adaptation samples. In detail, the accuracy interval achieved by our complete method ranges from 0.592 to 0.652 while MSN without FN ranges from 0.494 to 0.593. This figure reflects the viewer's individual difference and the influence of the meta-adaptation sample on the segmentation quality. For one viewer, the performances of MSNs are quite different due to different candidate meta-adaptation samples. However, since the results generated by BSN are the same for one viewer, the variances in the accuracies of the fused results become small. Furthermore, we calculate the average accuracy over 10 candidate meta-adaptation samples for each viewer, the IoUs of MSN without FN/our complete method are 0.525/0.593, 0.542/0.625, and 0.561/0.647. The comparisons also reveal that better results of MSN can possibly promote the accuracies of the final results. Examples of four groups of different viewers' salient object segmentation are shown in Fig. 9. We can see that MSN can learn different viewers' observation behaviors from the meta-adaptation samples accordingly. Although the segmentation results generated by MSN may have some noise, FN can improve the segmentation results by fusing the results generated by BSN so that our method can segment different viewers' salient objects according to their fixations' indications.

5.4. Ablation study

In order to analyze the effectiveness of proposed MSN's architecture, we performed an ablation study with three configurations: first one is to preserve the high branch and the low branch; second one is to preserve the high branch and the image branch; third one is to preserve the low branch and the image branch. These three configurations are illustrated in Fig. 10.

These three networks are re-trained by the same training procedure of the proposed MSN. Since the unsuitable transmission problem exists, the overall performance of all test samples cannot clearly demonstrate the performance of MSN. The analysis should focus on valid test samples which are indeed improved by MSN. According to the comparison of results generated by BSN and MSN, the result with higher accuracy using MSN can be treated as the valid sample. We count the number of the valid samples in the 3,000 test samples as same as those in the previous evaluation, and calculate their average IoU and average gained IoU as shown in Table 3. MSN without the image branch achieves the smallest number of the valid samples and the lowest accuracy. We can see that image features are very important which cannot be discarded. An improved segmentation result relies on the proper fusion of the features of image and observation behavior, which is learned by the ConvLSTM component. Although the number of the valid samples is the smallest, the high and the low branches guarantee this network to better extract the valid samples' observation behavior features so that it obtains higher gained IoU value. MSN without the low branch generates smaller number of the valid samples with lower accuracy and lower gained IoU value. The main reason is that the features of observation behavior extracted by the high branch are dependent on the fixation map

Fig. 7. Some segmentation results of our method. The fixations are indicated by the red dots. For one row, its columns from the left to the right represent one viewer's meta-adaptation sample with fixations, the test image with fixations, the ground truth, the result of BSN, the result of MSN and the result of FN.

only. It does not consider the influence of image content to observation behavior. The low branch focus on fixations' features dependent on the image. Therefore, many scenes with local similarity can benefit from the transmission by the low branch. However, this branch cannot well exploit the general features of the observation behavior. Therefore, MSN without the high branch can generate the most valid samples but the improvement is the smallest. Our proposed MSN utilizes the high branch and the low branch to comprise the features of the observation behavior. So, in the similar scenes, our proposed MSN can better learn observation behavior and achieves the largest improvement of the segmentation accuracy. It demonstrates that the network architecture of proposed MSN is reasonable.

5.5. Comparisons

Our method cooperates with the eye tracker device which can record the viewer's personalized fixation information. Thus, the meta -adaptation step performed for the viewer can be engaged in the calibration procedure when the eye tracker device is deployed. If segmented images are similar and can be prepared in advance, we can perform the meta adaptation step once by only one typical sample. Otherwise if images are diverse, in order to pursue the most suitable observation behavior for the transmission, we can collect several meta adaptation samples of each viewer. There are two possible strategies for segmentation. One is that we can run all samples of one viewer in one meta-

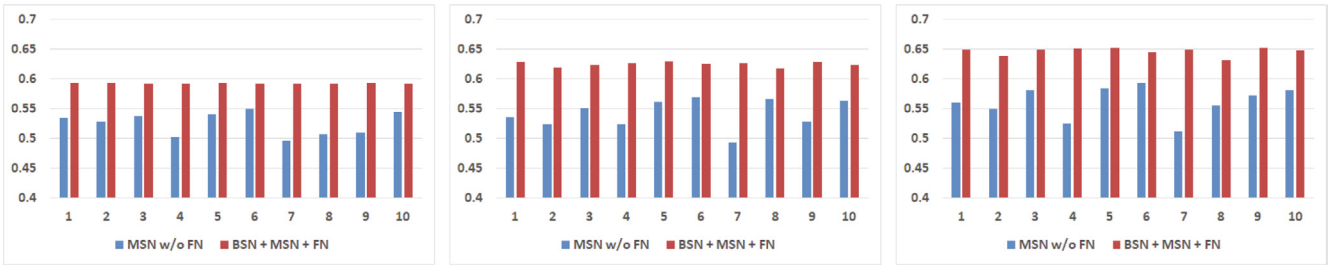


Fig. 8. The average accuracies of each viewer's 100 test samples using different meta-adaptation samples. (a), (b) and (c) represent three different viewers.

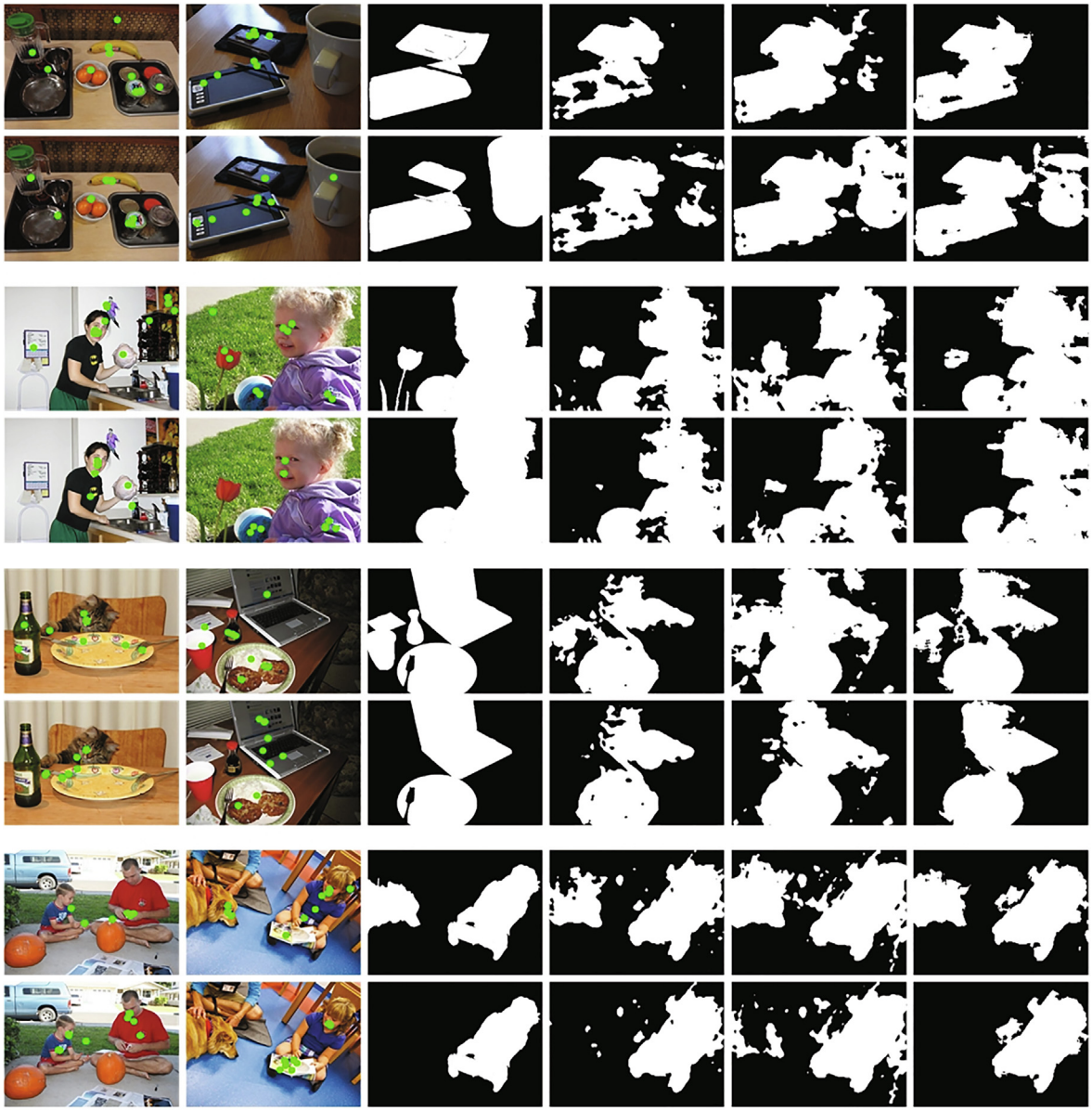


Fig. 9. Four groups of comparisons of different viewers' salient object segmentation. The fixations are indicated by the green dots. For one row, its columns from the left to the right represent one viewer's meta -adaptation sample with fixations, the test image with fixations, the ground truth, the result of BSN, the result of MSN and the result of FN.

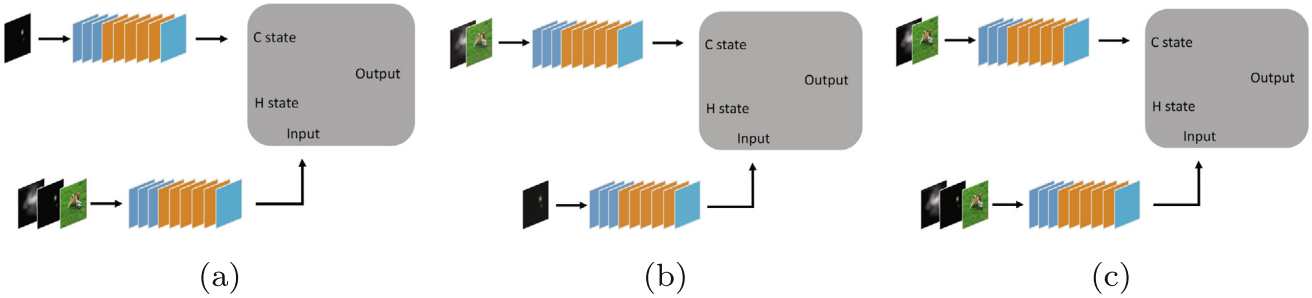


Fig. 10. Three network architectures in the ablation study. (a) MSN without the image branch (b) MSN without the low branch, (c) MSN without the high branch.

adaptation step (S1). Then, the MSN can segment the test image and fuse it with the result of the BSN by FN as the final result. The other strategy is that for each meta-adaptation sample, we perform one meta-adaptation step (S2). Thus, MSN can consequently generate multiple segmentation results of one test image. To record these possible results, we concatenate them and form a max-pooling result by selecting the maximum value of each pixel in the same position. Finally, the pooling result is fused with the

Table 3
The performances of four MSN in the ablation study.

Method	Number	*IoU* ↑	Gained *IoU* ↑
MSN w/o image branch	407	0.534	0.065
MSN w/o low branch	965	0.568	0.032
MSN w/o high branch	1227	0.580	0.021
MSN	1169	0.655	0.089

Table 4
Performances comparisons on OSIE-P test set.

Method	GBOS	SOS	UNet	CFPS
IoU ↑	0.375	0.411	0.586	0.631
Method	**Our S1**	**Our S1 + CFPS**	**Our S2**	**Our S2 + CFPS**
IoU ↑	0.624	0.646	0.622	0.649

result of BSN by FN as the final result. We tested these two strategies on the OSIE-P test set by 10 candidate meta -adaptation samples and 100 test samples as the same as those in the previous subsection for each viewer. Therefore, a total of 300 samples of three viewers are evaluated. We also compare our method against two other selection-aware methods, SOS [3] and GBOS [35]. In addition, we retrained CFPS [36] and the widely used segmentation network UNet [43] as other baselines, whose inputs are also composed of the image, the fixation map and the semantic map. In Table 4, we can see that our two strategies can achieve similar segmentation accuracies and their performances are better than those of other methods except for CFPS. The reason is that CFPS outperforms our simple BSN. However, if we use CFPS instead of BSN in the fusion stage, the results can be further improved by benefiting from the cooperation of MSN and CFPS. It indeed demonstrates that our MSN and FN are effective, and viewers' observation behavior transmission is a significant method to aid fixation based personalized salient object segmentation.

5.6. Limitation

As mentioned above, the unsuitable transmission problem heavily affects the performance of our segmentation method. Two failure cases induced by the unsuitable transmission problem are illustrated in Fig. 11. The upper case shows that the learned observation behavior for one dominant object is not suitable to be transmitted to the scene including separated small objects, and vice versa in the lower case. The unsuitable transmission of learned observation behavior makes MSN mistakenly estimate the sizes of salient objects. When results of BSN and MSN are both terrible, the effectiveness of FN is quite limited. Therefore, the optimization of the unsuitable transmission problem and the improvement of BSN are possible directions for our future work.

6. Conclusion

In this paper, we propose the fixation based personalized salient object segmentation method by utilizing the personalized observation behavior, where the personalized observation behavior is learned by the meta-learning. Besides the designed base segmentation network and the meta -segmentation network, we also develop the fusion network and further consider the weighted samples and the measure of the reliability of the meta-adaptation in the MAML to handle the unsuitable transmission of the observation behavior. Experimental results demonstrate that better segmentation performance of our method and the effectiveness of each network in our method.

CRediT authorship contribution statement

Ran Shi: Conceptualization, Methodology, Software, Writing - original draft, Writing - review & editing. **Gongyang Li:** Formal analysis, Validation. **Weijie Wei:** Data curation. **Xiaofei Zhou:** Supervision, Writing - review & editing. **Zhi Liu:** Writing - review & editing.

Declaration of Competing Interest

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

Acknowledgment

This work was supported by the National Natural Science Foundation of China under Grants 61801219, 61901145 and 61771301.

References

[1] Z. Chen, H. Zhou, J. Lai, L. Yang, X. Xie, Contour-aware loss: Boundary-aware learning for salient object segmentation, IEEE Transactions on Image Processing 30 (2021) 431–443, https://doi.org/10.1109/TIP.2020.3037536.
[2] L. Ye, Z. Liu, L. Li, L. Shen, C. Bai, Y. Wang, Salient object segmentation via effective integration of saliency and objectness, IEEE Transactions on Multimedia 19 (8) (2017) 1742–1756, https://doi.org/10.1109/TMM.2017.2693022.
[3] Y. Li, X. Hou, C. Koch, J.M. Rehg, A.L. Yuille, The secrets of salient object segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 280–287.
[4] A. Santella, M. Agrawala, D. DeCarlo, D. Salesin, M. Cohen, Gaze-based interaction for semi-automatic photo cropping, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, 2006, pp. 771–780.
[5] C. Guo, L. Zhang, A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression, IEEE Transactions on Image Processing 19 (1) (2010) 185–198, https://doi.org/10.1109/TIP.2009.2030969.
[6] M. Gygli, H. Grabner, H. Riemenschneider, F. Nater, L.V. Gool, The interestingness of images, in: IEEE International Conference on Computer Vision, 2013.

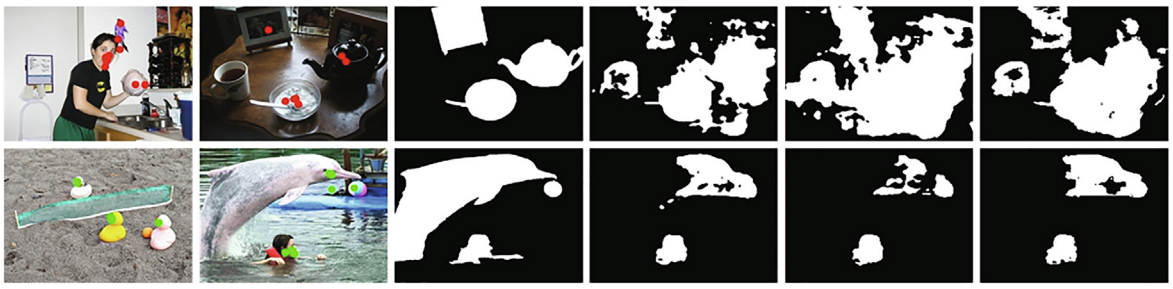


Fig. 11. Two examples of our failure cases. The fixations are indicated by the color dots. For one row, its columns from the left to the right represent one viewer's meta-adaptation sample with fixations, the test image with fixations, the ground truth, the result of BSN, the result of MSN and the result of FN.

[7] Y. Xu, N. Li, J. Wu, J. Yu, S. Gao, Beyond universal saliency: Personalized saliency prediction with multi-task cnn, in: Twenty-Sixth International Joint Conference on Artificial Intelligence, 2017. .
[8] M. Paul, M.M. Salehin, Spatial and motion saliency prediction method using eye tracker data for video summarization, IEEE Transactions on Circuits and Systems for Video Technology. .
[9] Y.Y. Boykov, M.-P. Jolly, Interactive graph cuts for optimal boundary & region segmentation of objects in nd images, in: Proceedings of the IEEE International Conference on Computer Vision, Vol. 1, IEEE, 2001, pp. 105–112. .
[10] M. Andrychowicz, M. Denil, S. Gomez, M.W. Hoffman, D. Pfau, T. Schaul, B. Shillingford, N. De Freitas, Learning to learn by gradient descent, arXiv preprint arXiv:1606.04474. .
[11] C. Finn, P. Abbeel, S. Levine, Model-agnostic meta-learning for fast adaptation of deep networks, 2017, arXiv:1703.03400. .
[12] Z. Li, F. Zhou, F. Chen, H. Li, Meta-sgd: Learning to learn quickly for few-shot learning, 2017, arXiv:1707.09835. .
[13] Y. Bengio, S. Bengio, J. Cloutier, Learning a synaptic learning rule, in: IJCNN-91-Seattle International Joint Conference on Neural Networks, Vol. ii, 1991, pp. 969 vol 2–. doi:10.1109/IJCNN.1991.155621. .
[14] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Computation 9 (8) (1997) 1735–1780.
[15] C. Chen, G. Wang, C. Peng, X. Zhang, H. Qin, Improved robust video saliency detection based on long-term spatial-temporal information, IEEE Transactions on Image Processing 29 (2020) 1090–1100, https://doi.org/10.1109/TIP.2019.2934350.
[16] G. Ma, C. Chen, S. Li, C. Peng, A. Hao, H. Qin, Salient object detection via multiple instance joint re-learning, IEEE Transactions on Multimedia 22 (2) (2020) 324–336, https://doi.org/10.1109/TMM.2019.2929943.
[17] W. Wang, H. Song, S. Zhao, J. Shen, S. Zhao, S.C.H. Hoi, H. Ling, Learning unsupervised video object segmentation through visual attention, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3059–3069, https://doi.org/10.1109/CVPR.2019.00318.
[18] Z. Yang, L. Huang, Y. Chen, Z. Wei, S. Ahn, G. Zelinsky, D. Samaras, M. Hoai, Predicting goal-directed human attention using inverse reinforcement learning, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 190–199, https://doi.org/10.1109/CVPR42600.2020.00027.
[19] G. Zelinsky, Z. Yang, L. Huang, Y. Chen, S. Ahn, Z. Wei, H. Adeli, D. Samaras, M. Hoai, Benchmarking gaze prediction for categorical visual search, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2019, pp. 828–836, https://doi.org/10.1109/CVPRW.2019.00111.
[20] L. Fan, Y. Chen, P. Wei, W. Wang, S. Zhu, Inferring shared attention in social scene videos, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 6460–6468, https://doi.org/10.1109/CVPR.2018.00676.
[21] A. Aydemir, K. Sjoo, J. Folkesson, A. Pronobis, P. Jensfelt, Search in the real world: Active visual object search based on spatial relations, in: 2011 IEEE International Conference on Robotics and Automation, 2011, pp. 2818–2824, https://doi.org/10.1109/ICRA.2011.5980495.
[22] A. Torralba, A. Oliva, M.S. Castelhano, J.M. Henderson, Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search, Psychological Review 113 (4) (2006) 766. .
[23] Z. Wei, H. Adeli, M.H. Nguyen, G. Zelinsky, D. Samaras, Learned region sparsity and diversity also predicts visual attention, in: Advances in Neural Information Processing Systems 29, 2016, pp. 1894–1902.
[24] H. Adeli, G. Zelinsky, Deep-bcn: Deep networks meet biased competition to create a brain-inspired model of attention control, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2018, pp. 2013–201310, https://doi.org/10.1109/CVPRW.2018.00259.
[25] N. Khosravan, H. Celik, B. Turkbey, R. Cheng, E. McCreedy, M. McAuliffe, S. Bednarova, E. Jones, X. Chen, P. Choyke, et al., Gaze2segment: a pilot study for integrating eye-tracking technology into medical image segmentation, in: Medical Computer Vision and Bayesian and Graphical Models for Biomedical Imaging, Springer, 2016, pp. 94–104.
[26] M. Sadeghi, G. Tien, G. Hamarneh, M.S. Atkins, Hands-free interactive image segmentation using eyegaze, in: Medical Imaging 2009: Computer-Aided Diagnosis, vol. 7260, International Society for Optics and Photonics, 2009, p. 72601H. .
[27] W. Wang, J. Shen, M.-M. Cheng, L. Shao, An iterative and cooperative top-down and bottom-up inference network for salient object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 5968–5977.
[28] W. Wang, J. Shen, Deep visual attention prediction, IEEE Transactions on Image Processing 27 (5) (2017) 2368–2378.
[29] W. Wang, J. Shen, J. Xie, M.-M. Cheng, H. Ling, A. Borji, Revisiting video saliency prediction in the deep learning era, IEEE Transactions on Pattern Analysis and Machine Intelligence 43 (1) (2019) 220–237.
[30] W. Wang, Q. Lai, H. Fu, J. Shen, H. Ling, R. Yang, Salient object detection in the deep learning era: An in-depth survey, IEEE Transactions on Pattern Analysis and Machine Intelligence. .
[31] W. Wang, J. Shen, X. Dong, A. Borji, R. Yang, Inferring salient objects from human fixations, IEEE Transactions on Pattern Analysis and Machine Intelligence 42 (8) (2020) 1913–1927, https://doi.org/10.1109/TPAMI.2019.2905607.
[32] W. Wang, J. Shen, X. Dong, A. Borji, Salient object detection driven by fixation prediction, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 1711–1720, https://doi.org/10.1109/CVPR.2018.00184.
[33] X. Tian, C. Jung, Point-cut, Fixation point-based image segmentation using random walk model, in: Proceedings of the IEEE International Conference on Image Processing, IEEE, 2015, pp. 2125–2129.
[34] S. Ramanathan, H. Katti, N. Sebe, M. Kankanhalli, T.-S. Chua, An eye fixation database for saliency detection in images, in: Computer Vision – ECCV 2010, Springer, Berlin Heidelberg, Berlin, Heidelberg, 2010, pp. 30–43. .
[35] R. Shi, N.K. Ngan, H. Li, Gaze-based object segmentation, IEEE Signal Processing Letters 24 (10) (2017) 1493–1497.
[36] G. Li, Z. Liu, R. Shi, W. Wei, Constrained fixation point based segmentation via deep neural network, Neurocomputing 368 (2019) 180–187.
[37] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556. .
[38] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440.
[39] L.-C. Chen, G. Papandreou, F. Schroff, H. Adam, Rethinking atrous convolution for semantic image segmentation, arXiv preprint arXiv:1706.05587. .
[40] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, W.-C. Woo, Convolutional lstm network: A machine learning approach for precipitation nowcasting, in: Advances in Neural Information Processing Systems, 2015, pp. 802–810. .
[41] J. Xu, M. Jiang, S. Wang, M.S. Kankanhalli, Q. Zhao, Predicting human gaze beyond pixels, Journal of Vision 14 (1) (2014) 28.
[42] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980. .
[43] O. Ronneberger, P.Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: Medical Image Computing and Computer-Assisted Intervention (MICCAI), vol. 9351 of LNCS, Springer, 2015, pp. 234–241. .

Ran Shi received his B.S. degree in Electronic Science and Technology from Changshu Institute of Technology and M.S. degree in Signal and Information Processing from Shanghai University in 2009 and 2012. He joined The Chinese University of Hong Kong (CUHK) as a Research Assistant in 2012, and obtained his Ph.D. in Electronic Engineering (CUHK) in 2017.
Currently, he is an assistant professor at the School of Computer Science and Engineering, Nanjing University of Science and Technology. His research interests include object segmentation, visual quality evaluation, interactive segmentation and salient object detection.

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

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

Xiaofei Zhou received the Ph.D. degree from Shanghai University, Shanghai, China, in 2018. He is currently a Lecturer with the School of Automation, Hangzhou Dianzi University. His research interests include saliency detection, and image/video segmentation.

Zhi Liu received the B.E. and M.E. degrees from Tianjin University, Tianjin, China, and the Ph.D. degree from Institute of Image Processing and Pattern Recognition, Shanghai Jiaotong University, Shanghai, China, in 1999, 20 02, and 20 05, respectively. He is currently a Professor with the School of Communication and Information Engineering, Shanghai University, Shanghai, China. From Aug. 2012 to Aug. 2014, he was a Visiting Researcher with the SIROCCO Team, IRISA/INRIA-Rennes, France, with the sup- port by EU FP7 Marie Curie Actions. He has published more than 170 refereed technical papers 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 in ICIP 2017, PCM 2016, VCIP 2016, ICME 2014, WIAMIS 2013, etc . He co-organized special sessions on visual attention, saliency models, and applications at WIAMIS 2013 and ICME 2014. He is an area editor of Signal Processing: Image Communication and served as a guest editor for the special issue on Recent Advances in Saliency Models, Applications and Evaluations in Signal Processing: Image Communication . He is a senior member of IEEE.