This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2018.2856126, IEEE Transactions on Multimedia

Saliency Integration: An Arbitrator Model

Yingyue Xu*, Xiaopeng Hong*, Fatih Porikli, *Fellow, IEEE*, Xin Liu, Jie Chen, and Guoying Zhao†, *Senior Member, IEEE*

***Abstract*—Saliency integration has attracted much attention on unifying saliency maps from multiple saliency models. Previous offline integration methods usually face two challenges: 1. if most of the candidate saliency models misjudge the saliency on an image, the integration result will lean heavily on those inferior candidate models; 2. an unawareness of the ground truth saliency labels brings difficulty in estimating the expertise of each candidate model. To address these problems, in this paper, we propose an arbitrator model (AM) for saliency integration. Firstly, we incorporate the consensus of multiple saliency models and the external knowledge into a reference map to effectively rectify the misleading by candidate models. Secondly, our quest for ways of estimating the expertise of the saliency models without ground truth labels gives rise to two distinct online model-expertise estimation methods. Finally, we derive a Bayesian integration framework to reconcile the saliency models of varying expertise and the reference map. To extensively evaluate the proposed AM model, we test twenty-seven state-of-the-art saliency models, covering both traditional and deep learning ones, on various combinations over four datasets. The evaluation results show that the AM model improves the performance substantially compared to the existing state-of-the-art integration methods, regardless of the chosen candidate saliency models.**

***Index Terms*—saliency integration, saliency aggregation, online model, arbitrator model.**

I. Introduction

Over the past two decades, saliency detection has hit much attention for its broad applications, such as image and video segmentation [1], video compression [2], and advertising [3]. Aiming at highlighting the regions of interest (ROI) of the human visual system on a scene with biologically plausible cues, a variety of saliency models have been proposed [4]–[26].

Existing saliency models utilize a broad range of strategies such as coarse-to-fine saliency map estimation [23], [27], top-down [10], [28] or bottom-up [4]–[8], [19], [29], [30] feature extraction, making different assumptions, for example, the background surrounding assumption [16], [18], [19], [31]–[33], and relying on a variety of models including support vector machine [10], AdaBoost [22], multiple kernel learning [11], [34], and deep convolutional neural networks [20], [21], [24], [35], [36], *etc*.

Recently, saliency integration (or saliency aggregation) approaches, which unify saliency maps from multiple existing saliency models, have attracted much attention [19], [37]–[41]. Although many of modern saliency models claim high performance in the statistical sense on different public benchmarks, none of them can outperform the others for every image under evaluation [42], [43]. For instance, even though the deep DHSNet [44] model, as one of the state-of-the-art approaches, is usually considered to surpass the traditional methods *e.g.*, GP [45], and MB+ [46], there are still images where DHSNet shows inferior predictions to GP and MB+, as shown in Figure 1. Thus, saliency (heat map) integration is proposed to take the advantages of multiple saliency models and make up for the defects of any specific ones, for enhanced accuracy and robustness of saliency detection.

Saliency integration is essentially a weighted combination of multiple saliency maps [37]. The weights, assuredly, play a central role in saliency integration. According to different ways of setting the weights, existing saliency integration approaches can be briefly categorized into the following two types.

Offline saliency integration models weigh candidate models by optimizing a specific energy function using a collection of data prepared in advance [37]–[40]. The integration model is fixed once the learning phase has been completed. However, they usually require extra efforts in providing the training samples with ground truth labels. Moreover, the scalability is limited [47], as the parameter settings of the integration model are only valid for a particular combination of candidate models. If a new candidate model is added, the integration model has to be retrained. Furthermore, there is an underline assumption that the *known* samples for learning and the *unknown* samples for prediction possess similar distributions. If the distribution of the *unknown* samples is significantly different from those of the *known* ones, the learnt parameters may fail in prediction.

Online integration models [19], [37], [41] are brought forth as a means of addressing the aforementioned problems of the *offline* models. *Online* models determine the weights of saliency maps by adapting to the image under evaluation directly, without the demand of any (pre-)collection of *known* samples. The resulting weights are, therefore, image-specific. Compared to the *offline* models, the *online* ones are free from fixing a model in advance and thus much more flexible and efficient. However, as every coin has two sides, online models have to face **two main challenges**.

1. **How to efficiently estimate the expertise of candidate models?** Most of the previous works assume that the expertise (*a.k.a.*, weights or contribution) of each candidate saliency model is equal (*e.g.*, BN [41] and MCA [19]). This assumption greatly eases the computational burden. However, it loses sight of the fact that each candidate saliency model shows discrepant ability in predicting an image. In fact, the performance of an integration without consideration of the expertise of candidate models may decrease, as the voices of the superior models

* The first two authors contributed equally. †Corresponding author.

Yingyue Xu, Xiaopeng Hong, Xin Liu, Jie Chen, and Guoying Zhao are with the Center for Machine Vision and Signal Analysis, University of Oulu, Finland. Emails: {yingyue.xu, xiaopeng.hong, xin.liu, jie.chen, guoying.zhao}@oulu.fi. Fatih Porikli is with the Research School of Engineering, Australian National University, Canberra. Email: fatih.porikli@anu.edu.au.

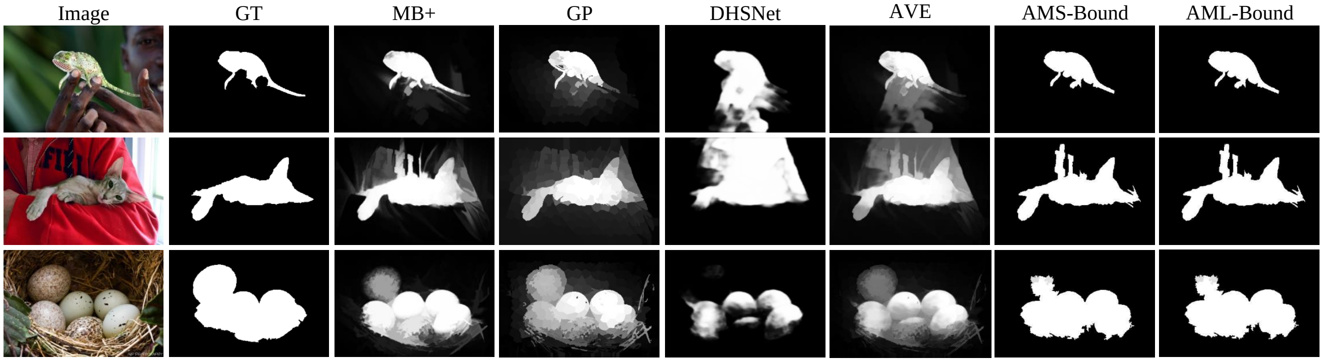


Fig. 1: Examples where saliency maps from the state-of-the-art deep model show inferior predictions to the traditional models. From left to right there are original images, ground truth (GT), traditional saliency models *e.g.*, MB+ [46] and GP [45], deep saliency model *e.g.*, DHSNet [44], naive integration approach *e.g.*, average map (AVE), and our proposed arbitrator model (AMS-Bound and AML-Bound). Examples are selected from the ECSSD dataset.

are easily drowned out by the mistakes made by those inferior ones. However, it is extremely difficult for online models to weigh each saliency model accurately, since there is no supervised information of the test images. Mai *et al.* [42] proposed to rank the performances of the saliency models on an image without the ground truth. However, since the ranking is a sequence of ordinal numbers, it cannot numerically measure the performance of each saliency model on the image in details. Le Meur *et al.* [37] estimate the expertise of the candidate saliency models by a weight function called M-estimator. The M-estimator decreases the expertise of the outliers that are detected according to their distances to a linear summation of the candidate saliency maps. However, as shown by the experimental results [37], the M-estimators perform similarly to *average weighting*, indicating that the computed weights are far from accurately specifying the expertise of the candidate models. Recently, some integration approaches [48], [49] explore expertise estimation by bringing the concept of superpixel difficulty, as each superpixel of an image may possess different difficulty for saliency assessment. This concept of using superpixel difficulty together with model expertise as hidden variables facilitates the expertise estimation process from a more refined superpixel level.

2. **How to ensure solid performance enhancement?** Le Meur *et al.* [37] also indicate that saliency integration models may decrease the performance in many cases. For instance, when most candidate saliency models misjudge a region on an image, the integration result will be highly susceptible to error. In Figure 2, we present the integration maps given by four typical online integration models using three popular saliency candidate methods on two images. The red rectangles on the ground truth indicate the regions that the candidate saliency models misjudge. From the integrated saliency maps, it can be perceived that when candidate saliency models misjudge a region on an image, the region will also be misjudged on the integrated map. Thus, overcoming the misleading by most of the candidate saliency models for solid performance enhancement becomes another big challenge in saliency integration.

This paper focuses on *online* saliency integration methods to treat the two challenges simultaneously:

(1) The saliency integration approach should efficiently determine the expertise of each candidate saliency model, in an online manner.

(2) The saliency integration method should have a mechanism to rectify the misleading by candidate models, even if most of the models misjudge a region on an image.

Based on the above two principles, we propose an online saliency integration framework, which is termed by the arbitrator model (AM) in this paper, as illustrated in Figure 3. We derive a Bayesian framework with the following two main components to reconcile the principles:

(1) We incorporate the consensus of multiple saliency models and the external knowledge into a reference map to effectively rectify the misleading by candidate models.

(2) Our quest for ways of estimating the expertise of the saliency models without ground truth labels gives rise to two distinct online model-expertise estimation methods: One is a statistical approach and the other is latent-variable-based. The two methods measure the expertise of the candidate saliency models without supervised information of the given test image, and meet the requirements of computing rational expertise of the candidate models.

The contributions of this paper are three-fold:

(1) We propose a Bayesian saliency integration framework, which makes the most of clues from both candidate saliency models and the external knowledge.

(2) We explore online ways of measuring the expertise of multiple saliency models and successfully introduce two methods.

(3) To extensively evaluate the proposed AM model, we test twenty-seven state-of-the-art candidate saliency models, including both traditional and deep learning ones, on various combinations over four datasets. To our best knowledge, the number of candidates and combinations under evaluation are the largest (ones) among the state-of-the-art saliency integration works.

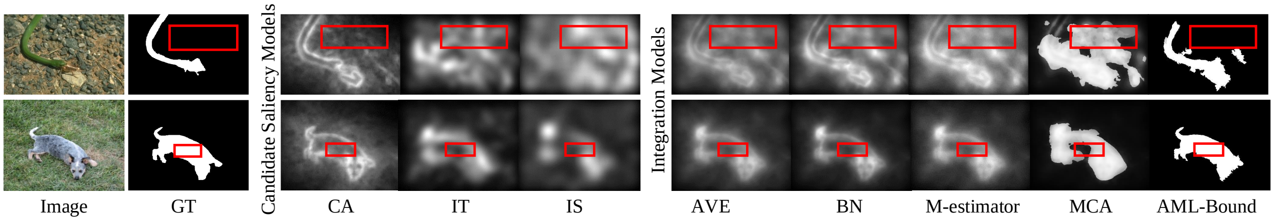


Fig. 2: Examples of misleading caused by misjudgement from candidate saliency models. From left to right columns are original images, ground truth (GT), candidate saliency models including CA [50], IT [4], IS [31], average map (AVE), integrated maps of BN [41], M-estimator [37], MCA [19], and our proposed arbitrator model (AML-Bound). The red rectangles on GT indicate the misjudged regions by the candidate saliency models.

II. Arbitrator Model

In this paper, we propose a Bayesian framework, namely the Arbitrator Model (AM), for saliency integration. As illustrated in Fig. 3, the AM model takes the test image and the corresponding saliency heat maps obtained by P saliency models as input. It consists of two main units: 1) a reference generator which makes use of the consensus of the input heat maps and the external knowledge; 2) an online estimator which treats the P saliency maps as candidates and evaluates their corresponding qualities (as termed by *expertise* hereinafter).

In the following of this section, we will derive the framework of the Arbitrator model (Section II-A) and provide an efficient solution for integration (Section II-B). The reference generator and the online estimator will be detailed in Sections III) and IV) respectively.

A. Bayesian Integration Framework

Superpixel algorithms group the pixels on an image into perceptually consistent units, and thus reserve the essential local structure of the image. It also efficiently reduces the computational costs of the subsequent processing tasks. Thus, the AM model proceeds to unify multiple saliency maps in the level of superpixel instead of the pixel level.

Given an image of N superpixels, each superpixel has a unique saliency label $l_n \in \{0,1\}$. We define the events that the n-th superpixel is salient (foreground) and inconspicuous (background) by F_n and $\bar{F}_n$ respectively. Apparently, we have $P\left(F_n\right) = P\left(l_n = 1\right)$, while $P\left(\bar{F_n}\right) = 1 - P\left(F_n\right) = P\left(l_n = 0\right)$.

Suppose there are P saliency models, each model is able to assign a saliency intensity value $s_{p,n} \in [0,1]$ to the n-th superpixel on the p-th saliency map. The binary saliency label of the n-th superpixel by the p-th model, is denoted as $\iota_{p,n} \in \{0,1\}$. $\iota_{p,n} = 1$ indicates the n-th superpixel is considered as a foreground one by the p-th model and vice versa. It can be easily obtained via a binarization process on the saliency intensity $s_{p,n}$ with a threshold γ_p, *e.g.*, OTSU thresholding [51]. More specifically, we have $\iota_{p,n} = 1$ ($\boldsymbol{\iota}_{p,n}$), if $s_{p,n} \geq \gamma_p$, otherwise, $\iota_{p,n} = 0$. Similarly, $\iota_{q\neq p,n} = 1$ ($\boldsymbol{\iota}_{q\neq p,n}$), if $s_{q\neq p,n} \geq \gamma_p$, otherwise, $\iota_{q\neq p,n} = 0$. Given the intensity of the n-th superpixel from the p-th model $s_{p,n}$ and the n-th superpixel being labeled as foreground on the binary maps by the rest models $\boldsymbol{\iota}_{q\neq p,n}$, the probability that the n-th superpixel is measured as foreground by the p-th model is $P\left(F_n|s_{p,n}, \boldsymbol{\iota}_{q\neq p,n}\right)$.[1]

The probability $P\left(F|s_p, \boldsymbol{\iota}_{q\neq p}\right)$ is derived under the Bayesian probability framework:

$$\begin{aligned} P\left(F|s_p, \boldsymbol{\iota}_{q\neq p}\right) &\propto P\left(F\right) P\left(s_p, \boldsymbol{\iota}_{q\neq p}|F\right) \\ &= P\left(F\right) P\left(s_p|F\right) P\left(\boldsymbol{\iota}_{q\neq p}|s_p, F\right) \\ &= P\left(F\right) P\left(s_p|F\right) \prod_{q\neq p} P\left(\boldsymbol{\iota}_q|F\right), \end{aligned} \tag{1}$$

with the assumption that all P saliency models make decisions independently, either with respect to the saliency intensity s or the binary saliency label ι. s_p represents the p-th saliency intensity map, while ι_p is the p-th binary saliency map.

The ratio of $\frac{P\left(F|s_p,\boldsymbol{\iota}_{q\neq p}\right)}{P\left(\bar{F}|s_p,\boldsymbol{\iota}_{q\neq p}\right)}$ is computed as follow:

$$\begin{aligned} \Lambda\left(F|s_p, \boldsymbol{\iota}_{q\neq p}\right) &= \frac{P\left(F|s_p, \boldsymbol{\iota}_{q\neq p}\right)}{P\left(\bar{F}|s_p, \boldsymbol{\iota}_{q\neq p}\right)} \\ &= \frac{P\left(F\right)}{P\left(\bar{F}\right)} \frac{P\left(s_p|F\right)}{P\left(s_p|\bar{F}\right)} \prod_{q\neq p} \frac{P\left(\boldsymbol{\iota}_q|F\right)}{P\left(\boldsymbol{\iota}_q|\bar{F}\right)} \end{aligned} \tag{2}$$

Then we compute the logarithm function of $\Lambda\left(F|s_p, \boldsymbol{\iota}_{q\neq p}\right)$ to form the integration framework, namely the arbitrator model (AM), as follow:

$$\begin{aligned} &\ln\Lambda\left(F|s_p, \boldsymbol{\iota}_{q\neq p}\right) \\ &= \ln\frac{P\left(F\right)}{P\left(\bar{F}\right)} + \ln\frac{P\left(s_p|F\right)}{P\left(s_p|\bar{F}\right)} + \sum_{q\neq p} \ln\frac{P\left(\boldsymbol{\iota}_q|F\right)}{P\left(\boldsymbol{\iota}_q|\bar{F}\right)} \end{aligned} \tag{3}$$

B. Implementation

We incorporate all the terms in Eq. 3 into a cellular automaton to generate a final saliency map.

Cellular Automaton (CA), *a.k.a.*, cellular space or homogeneous structure, is a discrete model in computability theory and mathematics [19], [52], [53]. A CA consists of a regular grid of cells. Each cell is with states, which are

[1] In this work, although the contexts with respect to F, $\bar{F}$, s and ι are upon a superpixel n, the sub-index n is omitted for clarity unless otherwise specified.

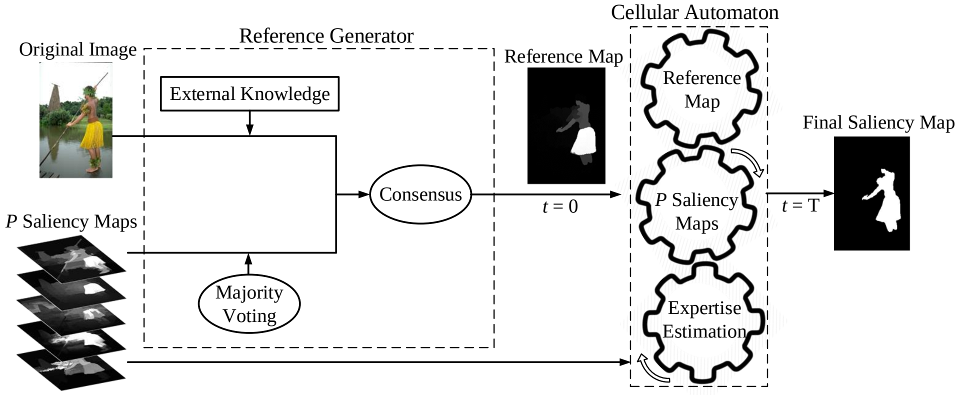


Fig. 3: Framework of the proposed arbitrator model (AM). The arbitrator incorporates the consensus of multiple saliency maps and the external knowledge into a reference map via the reference generator. A Bayesian integration framework reconciles the reference map and *P* saliency maps of varying expertise with cellular automaton (CA), to compute the final result. α_p and β_p are the expertise of the p-th saliency intensity map and the p-th binary map respectively. After each generation of the CA, the *P* saliency maps are updated. Accordingly, the expertise and the reference map are updated based on the new *P* saliency maps.

either discrete (*e.g.*, 'On' and 'Off') [52], [53] or continuous (*e.g.*, between 0 and 1) [19], [54]. The neighborhood of one specific cell can influence the states of the specific cell in next generations (advancing t by 1) in line with certain updating rules. Generally, the rule of updating the states of cells is a mathematical function, which is usually synchronous to all cells and time invariants.

Cellular automaton provides an efficient mechanism to propagate information to a batch of cells from their neighborhood respectively. For online saliency integration, we treat the fusion of the candidate saliency maps as a dynamic system by concentrating on the contextual relationships between different candidate saliency maps. More specifically, each unit (superpixels or pixels) of one candidate map is regarded as a cell and its corresponding saliency value as the 'state'. The units with the same locations on the other candidate maps become the neighbors of the cell. Considering the states within the neighborhood are contextually coherent, the states of the neighbors become indicators that orient one specific cell to evolve. One specific cell thus intends to make a wise decision for its state in the next generation, depending on the current states of its own and its neighbors. Intuitively, if its state is similar to the neighbors with high confidence, the cell should maintain its state; otherwise, the state should be updated towards the states of its high-confidence neighbors. As a result, as the CA updates itself iteratively, the states (saliency values) of all cells evolve for better estimation of the saliency values to make the dynamic system more stable.

As $\Lambda\left(F|s_p,\boldsymbol{\iota}_{q\neq p}\right) = \frac{P(F|s_p,\boldsymbol{\iota}_{q\neq p})}{P\left(\bar{F}|s_p,\boldsymbol{\iota}_{q\neq p}\right)} = \frac{P(F|s_p,\boldsymbol{\iota}_{q\neq p})}{1-P(F|s_p,\boldsymbol{\iota}_{q\neq p}))}$, the left side of Eq. 3 is the logarithm of the posterior ratio $\Lambda\left(F|s_p,\boldsymbol{\iota}_{q\neq p}\right)$ and thus a *logit* function of $P\left(F|s_p,\boldsymbol{\iota}_{q\neq p}\right)$. In this paper, $P\left(F|s_p,\boldsymbol{\iota}_{q\neq p}\right)$ is defined as s_p^{t+1}, which stands for the saliency value (of the n-th superpixel) on the p-th saliency intensity map at time $t+1$.

There are three terms on the right side of Eq. 3. 1) The first term $\ln\frac{P(F)}{P(\bar{F})} = \ln\frac{P(F)}{1-P(F)}$ is a logit function of $P(F)$. It is defined as logit(S_{Ref}^t), where S_{Ref}^t represents the saliency reference map at time t. The term at time 0 is initialized as the reference map S_{Ref}^0 which rectifies the misleading by candidate models. 2) The second term $\ln\frac{P(s_p|F)}{P\left(s_p|\bar{F}\right)}$ is the logarithm of the ratio of marginal likelihoods $\frac{P(s_p|F)}{P\left(s_p|\bar{F}\right)}$, which is clearly linked to the confidence or the reliability of the saliency intensity provided by the p-th candidate model. The second term is thus defined as $\ln\left(\alpha_p^t\right)\cdot s_p^t$, where α_p is the expertise of the p-th method and s_p^t is the p-th saliency intensity map at time t. 3) Similarly, the third term $\ln\frac{P(\boldsymbol{\iota}_q|F)}{P\left(\boldsymbol{\iota}_q|\bar{F}\right)}$ is associated with the confidence or the reliability of the q-th binary saliency map. Thus, we adopt $\mathcal{E}(s_q^t-\gamma_q^t)$ to threshold the q-th saliency map to obtain the corresponding binary one. Denoting the expertise by β_q, we define the third term as $\sum_{q\neq p}\ln\left(\beta_q^t\right)\cdot\mathcal{E}(s_q^t-\gamma_q^t)$.

At time $t=0$, the reference map S_{Ref}^0 is initialized from a reference generator as in Section III. At each generation $t>0$, we possess the reference map S_{Ref}^t and the saliency maps (s_p^t) of varying expertise α_p^t and β_p^t. Then we adopt CA to compute the corresponding S_{Ref}^{t+1}, s_p^{t+1}, α_p^{t+1} and β_p^{t+1} at time $t+1$. The synchronous updating rule of the cellular automaton derived from Eq. 3 is as follows:

$$\begin{aligned}\text{logit}\left(s_p^{t+1}\right) =&\text{logit}\left(S_{\text{Ref}}^t\right)\\ &+\ln\left(\alpha_p^t\right)\cdot s_p^t\\ &+\sum_{q\neq p}\ln\left(\beta_q^t\right)\cdot\mathcal{E}(s_q^t-\gamma_q^t).\end{aligned} \tag{4}$$

$$S_{\text{Ref}}^t=\frac{1}{P}\sum_{p=1}^{P}s_p^t. \tag{5}$$

$$S_{\text{Final}}=\frac{1}{P}\sum_{p=1}^{P}s_p^T. \tag{6}$$

The updating process of cellular automaton is illustrated in Figure 4. As empirically verified in experiments Section V, all the saliency intensity maps $s_p^T, p=1,\ldots,P$ will converge into stable states within 5 generations.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2018.2856126, IEEE Transactions on Multimedia

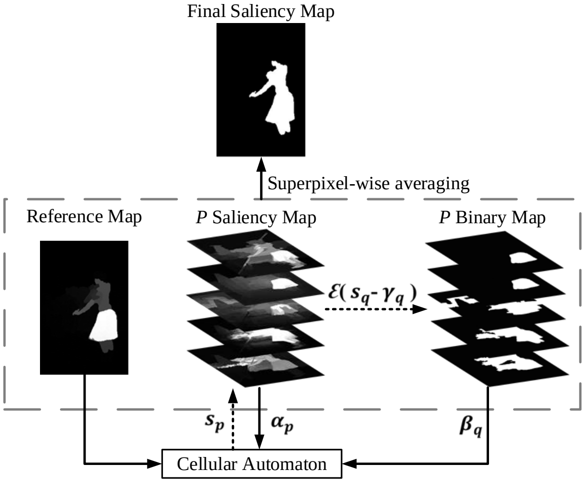


Fig. 4: Cellular automaton in the proposed arbitrator model. At each generation, the CA incorporates the reference map, the p-th saliency intensity map s_p with its expertise α_p and the rest binary maps with varying expertise β_q ($q \neq p$) by the expertise estimator into the final saliency map, as shown by solid arrows. s_p and the threshold γ_p are updated at each generation as shown by dashed arrows. $\mathcal{E}(s_q - \gamma_q)$ thresholds the q-th saliency intensity map to get the binary map.

CA has been adopted for saliency integration in MCA [19]. However, there is a strong assumption that the expertise of candidate models should be equal, which limits the integration performance as discussed in Section I. In our arbitrator model, we largely enrich the work by incorporating the reference map and candidate models with varying expertise into CA, which is out of the range of the previous MCA.

We will further introduce the computation of the reference map S^0_{Ref}, namely the reference generator, in Section III. Then we will detail the model-estimator that measures the expertise of saliency intensity maps α_p and the expertise of binary maps β_q in Section IV.

III. Reference Generator

As mentioned in Section I, there is possible misleading by the candidate saliency models. To overcome this problem for solid performance enhancement, we propose to hear voices from both the candidate models and the external knowledge. The reference map S^0_{Ref}, derived from $P(F)$ in Eq. 3, provides a natural scheme to introduce the external knowledge about salient object detection.

To acquire the reference map, we firstly compute an external saliency intensity map with external knowledge. Then, we compute a consensus map based on the consistency of the candidate models and the external knowledge map. Finally, we propagate the consensus map to get the reference map.

1) The External Knowledge Map: We introduce the external knowledge map to rectify the errors by the candidate models. Basically, the external knowledge can be any reasonable assumptions about salient object detection or currently existing saliency models. However the selection of the external knowledge is critical to the final integration performance.

In this work, we investigate three distinct methods to compute the external knowledge map. The first one is a handy and fast method based on the widely accepted assumptions, such as boundary prior [16], [19], [27], [55]–[60]. The second is the saliency map from one of the state-of-the-art traditional saliency models such as CCM [61]. The third one is the saliency map from one of the state-of-the-art deep models such as DHSNet model [44].

- **Knowledge Map from Assumptions**

In recent saliency detection approaches, it is widely accepted that the boundaries of a given image are most likely to be the background regions. Wei *et al.* [16], [55] point out that the most background regions, other than salient ones, are easily connected to image boundaries. This boundary prior theory comes from basic rules of photographic composition, and even if salient objects are far from the center, they seldom touch image boundaries (validated on MSRA datasetset in [55]). Similarly, a number of saliency models [19], [56], [57] generated a coarse saliency map with the compactness of image boundaries. Besides, several supervised saliency models [27], [58], [59] also extracted the appearance features of boundaries for model training. Therefore, we compute the external map S_{Ext} based on this boundary prior knowledge.

We assume that the more discrepant a superpixel is from the boundary superpixels, the more salient the superpixel is. We describe the mean CIELab feature of the n-th superpixel on an image as $\{c_n\}$ and select the superpixels along the four boundaries as boundary seeds. The boundary seeds are grouped into K clusters by K-means algorithm [62], c^k_b is the boundary superpixel belonging to the k-th cluster, K is empirically set as 3. For each superpixel c_n, we compute its appearance similarities to each cluster. If the superpixel is still quite different from its most similar cluster, it is more likely to be salient. In this way, we obtain the external knowledge map S_{Ext}:

$$S_{\text{Ext}}(c_n) = \min_{k\in\{1,\ldots,K\}} \left(\frac{1}{N_k}\sum_{b=1}^{N_k}\|c_n - c^k_b\|\right), \tag{7}$$

where N_k and $\|c_n - c^k_b\|$ are the number of superpixels in the k-th cluster, and the Euclidean distance between c_n and c^k_b.

- **Knowledge Map from Traditional Methods** We choose the saliency map from the contour-closure-based model (CCM) as one representative option for the external knowledge map. The CCM [61] highlights the importance of contour closure for salient object detection and propose to combine the closure completeness and the closure reliability for salient object detection. This CCM model proves to outperform the state-of-the-art unsupervised saliency models. Thus, the saliency map from CCM holds strong external knowledge to rectify inferior saliency information from traditional unsupervised candidate saliency models.

- **Knowledge Map from Deep Methods** If the deep saliency models are involved as candidates, the external knowledge based on traditional methods or assumptions will become relatively inferior. In such case, deep learning based external

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2018.2856126, IEEE Transactions on Multimedia

knowledge can be a preferred choice as the knowledge map. This paper introduces the resulted saliency map from the DHSNet model [44]. The DHSNet model utilizes a novel end-to-end deep hierarchical network based on convolutional neural networks for detecting salient objects. Evaluations prove that the DHSNet model shows significant superiority in terms of performance. Thus, we choose the saliency maps from DHSNet model as the external knowledge map if candidates involve deep models with extremely high performance.

2) Consensus: Even though an external knowledge map is introduced, its accuracy in saliency prediction can not be guaranteed just as the uncertainty in the candidate saliency models. Thus, we introduce a strict consistency scheme to reach a prudent consensus. In order to make a consensus, the arbitrator model judges the superpixel as salient only if the majority of the candidate saliency models vote it as salient as well as the external knowledge confirms its saliency. This consensus largely reduces the chances that an unsalient superpixel being misjudged as salient.

Given P saliency maps, $S_p(n)$ is defined as the mean intensity value of the n-th superpixel on the p-th saliency map. The majority voting map is computed as

$$S_{\text{Maj}}(n) = \begin{cases} 1, & \sum_{p=1}^{P} \iota_{p,n} > \frac{P}{2} \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

A consensus map S_{Con} is computed by hearing voices from both the majority voting map and the external knowledge map:

$$S_{\text{Con}} = S_{\text{Ext}} \times S_{\text{Maj}}, \tag{9}$$

The multiplication operation in Eq. 9 makes the consensus map S_{Con} constrained by the majority voting map S_{Maj} and the external knowledge map S_{Ext}. It largely reduces the numbers of false positives on these two maps.

3) The Reference Map via Propagation: The consensus map S_{Con} is a saliency map of high precision but only holds saliency information for certain parts of the image, so that we need to expand the saliency information to the whole image. A propagation method is employed to diffuse these saliency values on the consensus map to the whole image iteratively.

Propagation method firstly over-segments the image into superpixels and constructs an undirected graph, which comprises of a set of vertices of the superpixels together with a set of edges representing the similarity between adjacent vertices. Then, the propagation seeds [23], [63]–[66] are selected to spatially diffuse to the whole graph within several iterations.

The over-segmented image can be regarded as an undirected graph $G = (V, E)$, which comprises a set V of the superpixels together with a set E of edges representing the similarity between adjacent superpixels. The constructed graph G can be described as an adjacent matrix $W = [w_{nm}]_{N\times N}$ with the similarity between two superpixels c_n and c_m computed as $w_{nm} = exp(-\mathcal{G}(c_n, c_m)^2/(2\theta^2))$, where θ is set as 0.25 and $\mathcal{G}(c_n, c_m)$ computes the geodesic distance between c_n and c_m:

$$\mathcal{G}(c_n, c_m) = \min_{V_1=n, V_2, \dots, V_r=m} \left[\sum_{k=1}^{r-1} \max(\|V_k - V_{k+1}\| - a, 0) \right] \tag{10}$$

s.t. $V_k, V_{k+1} \in V$, $\|V_k - V_{k+1}\|$ computes the Euclidean distance between V_k and V_{k+1}, and a is an adaptive threshold preventing the "small-weight-accumulation" problem [23], [55]. The $\mathcal{G}(c_n, c_m)$ measures the shortest path between c_n and c_m in the graph G.

Finally, we use a propagation function as follow to compute the reference map:

$$S^{t+1} = I \cdot D^{-1} \cdot W \cdot S^t. \tag{11}$$

where I is the identity matrix and D is the diagonal degree matrix with $D_{nm} = \sum_m w_{nm}$, the initial $S^{t=0} = S_{\text{Con}}$, and after several times of iterations, the final propagated map is computed as the reference map S^0_{Ref}. In practice, we set the propagation number as 5.

In this section, we propagate a reference map by taking the consensus of the external knowledge and the majority voting of all the candidate saliency models into consideration. The reference map integrated with the external knowledge is regarded as the reference map S^0_{Ref}. Afterwards, each candidate saliency map is updated based on Eq. 4. Accordingly, the reference map S^t_{Ref} at $t > 0$ is updated by averaging the candidate saliency maps as in Eq. 5. Thus, during the CA updating process, the reference map is updated by using the candidate saliency maps as in Eq. 5 and the external knowledge is not integrated any more. The influence of the external knowledge is thus appropriate and recessive.

IV. Model-expertise Estimator

The α_p and β_p measure the expertise of the p-th saliency model. α_p represents the expertise of the p-th saliency intensity map, which is a map with continuous values in the range of $[0, 1]$. β_p represents the expertise of the p-th binary map, which is a map with binary values $\{0, 1\}$.

The saliency maps integrated into the framework involve both the intensity maps and the binary maps. In this work, the cellular automata (CA) is used to iteratively update the current saliency intensity map by involving the contextual influences of its neighborhood (other candidate saliency maps). Firstly, it is natural to design the candidate saliency map as an intensity map of continuous values for finer estimation, as it reflects the actual saliency intensity by the corresponding candidate saliency model. However, as intensity maps from various saliency models have diverse semantic meanings and magnitudes, we also adopt their corresponding binary maps as the contextual neighborhood to eliminate these two influences. It is convenient to transform an intensity map to a binary saliency map when necessary by introducing a threshold (like Otsu [45]). Otherwise, if it was the binary map to be updated in CA, it would cause inevitably information loss at the very beginning of the integration, and it would be impossible to recover an intensity one. Therefore, in the framework, both saliency intensity maps and saliency binary maps are integrated, but the current saliency maps to be updated in CA are intensity maps.

In this paper, we propose two online approaches to obtain the expertise of saliency models without supervised information from the image dataset, one is a statistics-based method from the intrinsic implications of Eq. 3, the other is a latent-variable-based method for evaluating multiple models.

A. Statistics-based Expertise

According to the Bayesian framework of Eq. 3, we propose a statistics-based method to evaluate the expertise α_p and β_p. The statistics-based method analyzes the probability distributions of foreground and background samples on saliency maps and statistically computes α_p and β_p.

β_p is the expertise of the p-th binary saliency map, which is originally derived from $\frac{P(\boldsymbol{\iota}_p|F)}{P(\boldsymbol{\iota}_p|\bar{F})}$. More specifically, $P(\boldsymbol{\iota}_p|F)$ is $P(\iota_{p,n}=1|F_n)$, indicating the probability that the n-th superpixel on the p-th saliency map is labeled as foreground given the superpixel is a foreground one. Similarly, $P(\boldsymbol{\iota}_p|\bar{F})$ is $P(\iota_{p,n}=1|\bar{F}_n)$, indicating the probability that the n-th superpixel is miss-labeled as foreground given the superpixel is a background one. Although it is impossible to get the ground-truth F in online methods, the reference map obtained in Section III can be regarded as the 'best current' knowledge to approximate F.

In this work, as the burden of computing every local $\beta_{p,n}$ is rather heavy, we set a threshold λ to classify the reference map $P(F)$ as foreground or background samples, and estimate a global β_p to approximate the expertise of all the superpixels on the p-th saliency map. Then, the computation of β_p is simplified as follows:

$$\beta_p = \frac{P(\boldsymbol{\iota}_p|F)}{P(\boldsymbol{\iota}_p|\bar{F})} = \frac{P(\iota_{p,n}=1|F_n)}{P(\iota_{p,n}=1|\bar{F}_n)} \propto \frac{P(\iota_p=1|F)}{P(\iota_p=1|\bar{F})} \tag{12}$$

Thus, $P(\iota_p=1|F)$ and $P(\iota_p=1|\bar{F})$ can be obtained only by their intrinsic implications of probability theory, namely

$$P(\iota_p=1|F) = \frac{P(\iota_p=1,F)}{P(F)}, \tag{13}$$

$$P(\iota_p=1|\bar{F}) = \frac{P(\iota_p=1,\bar{F})}{P(\bar{F})}, \tag{14}$$

More specifically, the probability functions $P(\iota_p=1,F)$ and $P(\iota_p=1,\bar{F})$ are statistically computed as follow:

$$P(\iota_p=1,F) = \frac{1}{N}\sum_{n=1}^{N}[\mathcal{T}(S_p(n),\gamma_p)\cdot\mathcal{T}(S_{\text{Ref}}(n),\lambda)], \tag{15}$$

$$P(\iota_p=1,\bar{F}) = \frac{1}{N}\sum_{n=1}^{N}[\mathcal{T}(S_p(n),\gamma_p)\cdot(1-\mathcal{T}(S_{\text{Ref}}(n),\lambda))], \tag{16}$$

where N is the number of superpixels on the over-segmented image. $S_p(n)$ is the mean intensity value of the n-th superpixel on the p-th saliency intensity map, and γ_p is the OTSU threshold [51] of the p-th saliency intensity map. $S_{\text{Ref}}(n)$ is the mean intensity value of the n-th superpixel of the reference map. $\mathcal{T}$ is a thresholding function as follow:

$$\mathcal{T}(\mu,\nu) = \begin{cases} 1, & \mu \geq \nu \\ 0, & \text{otherwise}, \end{cases} \tag{17}$$

$P(F)$ is computed as

$$P(F) = \frac{1}{N}\sum_{n=1}^{N}\mathcal{T}(S_{\text{Ref}}(n),\lambda), \tag{18}$$

and $P(\bar{F}) = 1 - P(F)$. Thus, a global β_p of the p-th binary saliency map is computed based on probability theory.

α_p represents the expertise of the p-th saliency intensity map, which can be computed in a similar way as computing β_p. Thus, we have

$$\alpha_p = \frac{P(s_p|F)}{P(s_p|\bar{F})} = \frac{\frac{P(s_p,F)}{P(F)}}{\frac{P(s_p,\bar{F})}{P(\bar{F})}} \tag{19}$$

However, as s_p is a map with continuous values other than discrete ones, we employ a fixed stepsize of 0.1 in $[0.1, 0.9]$ to binarize s_p with gradually increasing thresholds, and α_p is the mean ratio of $\frac{P(\boldsymbol{\iota}_p|F)}{P(\boldsymbol{\iota}_p|\bar{F})}$ with all the thresholds.

We denote the step-sized thresholds as a set of $\lambda^{'} = \{0.1, 0.2, 0.3, \ldots, 0.9\}$, where $J=9$ is the number of thresholds in the set $\lambda^{'}$. Then the probability functions $P(s_p,F)$ and $P(s_p,\bar{F})$ are statistically computed as

$$\begin{aligned} &P(s_p,F) \\ &= \frac{1}{N}\frac{1}{J}\sum_{j=1}^{J}\sum_{n=1}^{N}[\mathcal{T}(S_p(n),\gamma_p)\cdot\mathcal{T}(S_{\text{Ref}}(n),\lambda^{'}(j))], \end{aligned} \tag{20}$$

$$\begin{aligned} &P(s_p,\bar{F}) \\ &= \frac{1}{N}\frac{1}{J}\sum_{j=1}^{J}\sum_{n=1}^{N}[\mathcal{T}(S_p(n),\gamma_p)\cdot(1-\mathcal{T}(S_{\text{Ref}}(n),\lambda^{'}(j)))]. \end{aligned} \tag{21}$$

With the probability theory, we finally compute the expertise of the p-th saliency intensity map α_p and the expertise of the p-th binary saliency map β_p in a statistical way.

B. Latent-variable-based Expertise

The candidate saliency models distinguish a superpixel on an image as salient or not based on the corresponding saliency maps. Besides, each superpixel on the image is assumed to possess difficulty for saliency assessment, namely π_n. In recent saliency integration approaches [48], [49], the concept of superpixel difficulty are adopted in the process of computing the expertise of the candidate saliency map. The expertise β_p as well as the difficulty of the superpixel π_n are assumed as latent variables and are solved by optimizations.

β_p represents the expertise of the p-th binary saliency map, which is assumed to range $\beta_p \in (-\infty, +\infty)$. If $\beta_p < 0$, the p-th candidate model makes wrong measurements and shows inferior ability in saliency prediction. If $\beta_p > 0$, the p-th model makes correct measurements and shows superior ability in saliency prediction. When $\beta_p = 0$, the p-th model is not able to distinguish saliency objects. $\beta_p = +\infty$ implicates that the p-th model always makes correct decisions about saliency objects, while $\beta_p = -\infty$ means that the p-th binary saliency map always misjudge saliency information.

Model	DRFI	MB+	RB	TLLT	MB	BSCA	RC	MR	GP	UFO	COV
MAE	0.170	0.171	0.172	0.172	0.174	0.183	0.186	0.186	0.191	0.203	0.220
F-measure	0.765	0.722	0.710	0.717	0.709	0.736	0.720	0.735	0.725	0.684	0.602

Model	HS	GC	CEOS	PCAS	GBVS	LR	IT	FT	CA	SR	IS
MAE	0.228	0.234	0.243	0.247	0.263	0.274	0.289	0.291	0.309	0.311	0.334
F-measure	0.700	0.564	0.646	0.622	0.599	0.622	0.520	0.384	0.483	0.409	0.416

Model	DHSNet	DSS	DCL	MDF	RFCN	CCM	Bound
MAE	0.059	0.062	0.068	0.135	0.147	0.151	0.237
F-measure	0.886	0.884	0.888	0.796	0.878	0.738	0.581

TABLE I: The list of the twenty-seven candidate saliency models. The models are ranked by their MAE evaluation results on the ECSSD dataset and their mean F-measure scores (with adaptive thresholds) are also reported. In the bottom part, the performances of deep models, CCM model and the pre-computed boundary-prior-based external knowledge map are presented.

Besides the assumption that candidate models vary in expertise β_p, we presume that each superpixel in an image has varying degrees of difficulty for saliency assessment and introduce a measurement $\pi_n \in [0, +\infty)$ to represent the difficulty of a superpixel. $\pi_n = 0$ means that the superpixel possesses extremely low difficulty such that even an inexperienced saliency model can distinguish its saliency. On the contrary, $\pi_n = +\infty$ means the superpixel is so ambiguous that even the best saliency model has a chance to misjudge it.

As defined in Section II-A, l_n is the true binary saliency label of the n-th superpixel on the given image, while $\iota_{p,n}$ refers to the actual binary saliency label of the n-th superpixel by the p-th model. Thus, the probability that the p-th model correctly labels a superpixel on an image is

$$p\left(\iota_{p,n} = l_n | \beta_p, \pi_n\right) = \begin{cases} 1, & \pi_n = 0 \\ \frac{1}{1+e^{-\beta_p/\pi_n}}, & \text{otherwise} \end{cases} \tag{22}$$

More skilled saliency models (higher β_p) have a higher probability of correctly labeling a superpixel. As the difficulty π_n of a superpixel increases, the probability of correctly labeling the superpixel decreases, and vice versa.

Now, given a set of actual saliency labels by multiple saliency models $\boldsymbol{\iota} = \{\iota_{p,n}\}$, our goal is to estimate the unobserved latent parameters including the true saliency labels of superpixels $\boldsymbol{l} = \{l_n\}$, the expertise of the candidate models $\boldsymbol{\beta} = \{\beta_p\}$, and the difficulties of the superpixels $\boldsymbol{\pi} = \{\pi_n\}$. Here the Expectation-Maximization (EM) algorithm is used to achieve the optimal values of the latent parameters.

In the E-step, we compute the posterior probabilities of l_n with the parameters $\boldsymbol{\beta}$, $\boldsymbol{\pi}$ obtained from the last M-step and the actual labels:

$$\begin{aligned} p(l_n|\boldsymbol{\iota}, \boldsymbol{\beta}, \boldsymbol{\pi}) &= p(l_n|\boldsymbol{\iota}_n, \boldsymbol{\beta}, \pi_n) \\ &\propto p(l_n|\boldsymbol{\beta}, \pi_n) p(\boldsymbol{\iota}_n|l_n, \boldsymbol{\beta}, \pi_n) \\ &\propto p(l_n) \prod_p (\iota_{p,n}|l_n, \beta_p, \pi_n), \end{aligned} \tag{23}$$

where $\boldsymbol{\iota}_n$ denotes the actual labels of a superpixel by all the P candidate models and the parameters $\boldsymbol{\beta}$, $\boldsymbol{\pi}$ are conditionally independent of l_n. In practice, we use Gaussian distribution ($\mu = \theta = 1$) for β, re-sample $1/\boldsymbol{\pi}$ as $e^{(1/\boldsymbol{\pi}')}$, and use the same Gaussian distribution on $1/\boldsymbol{\pi}'$ to avoid $\boldsymbol{\pi}$ being negative.

In the M-step, we compute the expected value of the log likelihood function with respect to the conditional distribution of $\boldsymbol{l}$ given $\boldsymbol{\iota}$ under the current estimate of $\boldsymbol{\beta}$ and $\boldsymbol{\pi}$ as follows:

$$\begin{aligned} Q\left(\boldsymbol{\beta}, \boldsymbol{\pi}\right) &= E\left[\ln p\left(\boldsymbol{\iota}, \mathbf{l} | \boldsymbol{\beta}, \boldsymbol{\pi}\right)\right] \\ &= E\left[\ln \prod_n \left(p\left(l_n\right) \prod_p p\left(\iota_{p,n} | l_n, \beta_p, \pi_n\right)\right)\right] \\ &= \sum_n E\left[\ln p\left(l_n\right)\right] + \sum_{p,n} E\left[\ln p\left(\iota_{p,n} | l_n, \beta_p, \pi_n\right)\right], \end{aligned} \tag{24}$$

since $\iota_{p,n}$ are conditionally independent given $\boldsymbol{l}$, $\boldsymbol{\beta}$, $\boldsymbol{\pi}$. With gradient ascent method, the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\pi}$ are set to maximize the quantity function Q in Eq. 24.

Here, we presume that the expertise of the p-th saliency intensity map α_p is equal to β_p to simplify the computation. The details of the EM algorithm can be found in [67].

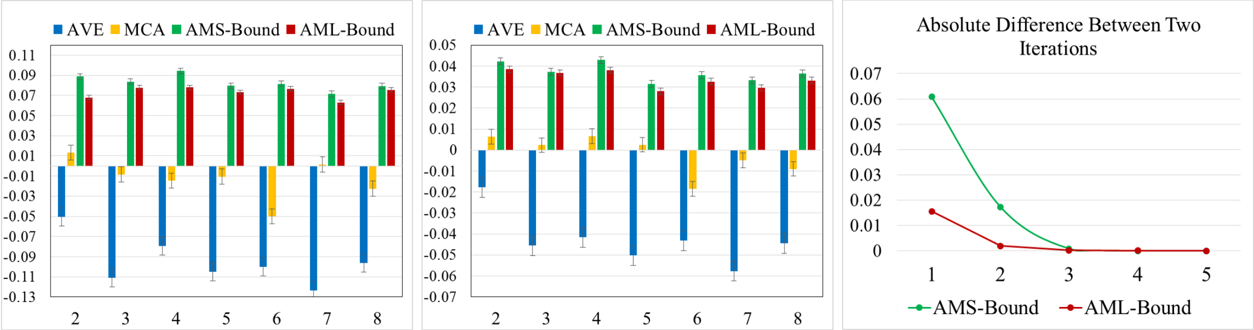
V. Experiments

The arbitrator model (AM) aims at generating a saliency integration model that solidly enhances the performance regardless of the choices of candidate saliency models. Any saliency models can be selected for saliency integration in AM and no special assumptions on saliency models are required.

In this section, we perform a comprehensive evaluation of the AM model under various combination strategies by adopting the state-of-the-art saliency models as the candidates. We choose twenty-seven state-of-the-art saliency models including the traditional models BSCA [19], CA [50], CEOS [68], COV [29], DRFI [58], FT [9], GBVS [5], GC [14], GP [45], HS [69], IS [31], IT [4], LR [32], MB [46], MB+ [46], MR [66], PCAS [30], RB [55], RC [14], SR [8], TLLT [23], UFO [58], and deep models including DSS [70], DCL [71], RFCN [72], MDF [73], and DHSNet [44]. The implementations of the chosen approaches are directly from the corresponding authors.

For comprehensive evaluation, four challenging datasets are utilized in the experiments: ECSSD [69], ASD [9], ImgSal [74] and DUT-OMRON [66]. The ASD dataset is one of the most widely used datasets with 1000 images from the MSRA-5000 Saliency Object Database [75], with distinct salient objects on the scenes. The ImgSal dataset is challenging, including 235 images in six levels of complexity. The ECSSD dataset contains 1000 images with complex salient

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2018.2856126, IEEE Transactions on Multimedia



(a) F-measure (b) MAE (c) Convergence Experiment

Fig. 5: (a)-(b) The average performance enhancement of five randomly selected combinations for each fixed number combination using strategy 3 in Section V-B compared to their corresponding top models. (a) measures the mean F-measure improvement. (b) measures the average improvement of MAE scores. The average maps (AVE), MCA, AMS and AML results are compared and horizontal axis indicates the number of candidate models being combined. (c) Convergence experiments computing the average absolute difference of all superpixels between two generations.

objects on the scenes, and the objects on the images are semantically meaningful. The DUT-OMRON dataset contains a large number of 5168 more difficult and challenging images.

A. Implementation and Evaluation

We over-segment the images into $N = 400$ superpixels with the simple linear iterative clustering (SLIC) algorithm [76]. In practice, we set the numbers of generations in Eq. 6 as 5 for the CA updates, and λ is set as 0.1. We reference our arbitrator model with statistics-based expertise as AMS and with latent-variable-based expertise as AML in all the experiments. Moreover, we refer "-B", "-C", "-D" as the boundary-based external knowledge, contour-closure-based external knowledge and deep-based external knowledge respectively (*i.e.*, AMS-B means that the AM model uses the boundary-based external knowledge in the reference generator and adopts the statistics-based expertise estimator). Besides the saliency maps computed from AMS and AML, we also compute the average saliency maps of the candidate saliency models (AVE), BN [41], M-estimator [37] and MCA [19] for fair comparisons. All the existing saliency integration models being selected in this work are online models, and the codes are provided by the corresponding authors with recommended parameter settings.

We employ two types of evaluation metrics to evaluate the performance of saliency maps: F-measure and mean absolute error (MAE). F-measure is computed to count for the saliency maps with both high precision and recall:

$$F = \frac{\left(1+\beta^2\right) \cdot precision \cdot recall}{\beta^2 \cdot precision + recall}, \tag{25}$$

where $\beta^2 = 0.3$ [9] to emphasize the precision, and the precision and recall are obtained by using twice the mean saliency values of the saliency maps as adaptive thresholds [9].

MAE measures the overall pixel-wise difference between the saliency map sal and the ground truth gt:

$$MAE = \frac{1}{H} \sum_{h=1}^{H} |sal(h) - gt(h)|. \tag{26}$$

where H is the number of pixels on the map.

B. Comparisons of Various Combinations

We choose the number of candidate saliency models for integration from 2 to 8. If enumerating all the possible combinations from 2 to 8 models, we need to evaluate $C_{27}^2 + C_{27}^3 \ldots + C_{27}^8 = 3,505,671$ combinations, which is almost impossible. Thus, we follow four different strategies to evaluate fifty-eight representative combinations. Table I lists the performances of the twenty-seven candidate saliency models on ECSSD dataset by ranking the MAE.

In Table II, we list the mean F-measure of the proposed AM model with four different combination strategies. We compare the integrated results of the AM model with every candidate saliency model being combined (only list the one with the best performance in column "Top" and refer it as top model in experimental analysis), the average saliency maps (AVE), the resulted BN, M-estimator (M-est), and MCA saliency maps. The detailed evaluation and analysis of the four strategies are listed below.

1. **Superior models combination.** When choosing the candidate saliency models, we only consider those saliency models with the best performances. Thus, we choose two best saliency models for 2-model-combination, three best saliency models for 3-model-combination and so forth. The first seven rows in Table II indicate the evaluation results, where it can be easily perceived that both AMS and AML outperform the top candidate saliency model as well as the MCA model in every combination. Thus, the proposed AM model performs well when superior saliency models are combined.

2. **Inferior models combination.** We only consider those saliency models with the worst performances. For example, we choose the worst two saliency models for 2-model-combination, the worst three saliency models for 3-model-

Combination		Top	AVE	BN	MCA	M-est	AMS-B	AMS-C	AMS-D	AML-B	AML-C	AML-D
Superior Models Combinations	DRFI, MB+	0.765	0.750	0.750	0.760	0.752	0.791	0.794	**0.832**	0.785	0.803	**0.850**
	DRFI, MB+, RB	0.765	0.752	0.752	0.765	0.754	0.772	0.780	**0.827**	0.772	0.788	**0.851**
	DRFI, MB+, RB, TLLT	0.765	0.762	0.762	0.765	0.764	0.773	0.784	**0.821**	0.766	0.781	**0.830**
	DRFI, MB+, RB, TLLT MB	0.765	0.753	0.753	0.758	0.755	0.770	0.775	**0.819**	0.768	0.780	**0.836**
	DRFI, MB+, RB, TLLT MB, BSCA	0.765	0.757	0.757	0.762	0.759	0.778	0.781	**0.816**	0.774	0.784	**0.832**
	DRFI, MB+, RB, TLLT MB, BSCA, RC	0.765	0.764	0.764	0.767	0.766	0.783	0.783	**0.821**	0.778	0.787	**0.837**
	DRFI, MB+, RB, TLLT MB, BSCA, RC, MR	0.765	0.765	0.765	0.767	0.768	0.783	0.785	**0.815**	0.781	0.788	**0.828**
Inferior Models Combinations	SR, IS	0.416	0.420	0.420	0.442	0.420	0.556	0.613	**0.688**	0.576	0.637	**0.723**
	CA, SR, IS	0.483	0.456	0.456	0.472	0.456	0.554	0.609	**0.690**	0.579	0.647	**0.738**
	FT, CA, SR, IS	0.483	0.465	0.465	0.491	0.465	0.550	0.612	**0.677**	0.567	0.635	**0.715**
	FT, CA, SR, IS IT	0.520	0.492	0.492	0.513	0.492	0.585	0.636	**0.711**	0.602	0.658	**0.744**
	FT, CA, SR, IS LR, IT	0.622	0.537	0.537	0.554	0.537	0.632	0.670	**0.734**	0.627	0.674	**0.750**
	FT, CA, SR, IS GBVS, LR, IT	0.622	0.562	0.562	0.574	0.562	0.669	0.692	**0.756**	0.662	0.702	**0.783**
	FT, CA, SR, IS PCAS, GBVS, LR, IT	0.622	0.583	0.583	0.595	0.583	0.685	0.703	**0.755**	0.675	0.707	**0.778**
Random Combinations	DRFI, GP	0.765	0.768	0.768	0.776	0.770	0.793	0.794	**0.826**	0.789	0.803	**0.846**
	HS, IT, IS	0.700	0.636	0.636	0.592	0.633	0.708	0.719	**0.776**	0.705	0.745	**0.835**
	HS, GC, COV, CA	0.700	0.647	0.647	0.585	0.647	0.701	0.719	**0.765**	0.709	0.715	**0.779**
	HS, GC, COV, CA MR	0.735	0.727	0.728	0.723	0.729	0.736	0.755	**0.796**	0.735	0.772	**0.828**
	MB, BSCA, RC, GBVS COV, FT	0.736	0.733	0.734	0.743	0.734	0.744	0.755	**0.790**	0.747	0.770	**0.825**
	MB+, GP, BSCA, RB GBVS, IT, IS	0.736	0.736	0.736	0.726	0.736	0.751	0.757	**0.794**	0.758	0.764	**0.826**
	MB, BSCA, TLLT, GC PCAS, GBVS, HS, CA	0.736	0.742	0.743	0.724	0.743	0.761	0.768	**0.801**	0.755	0.770	**0.822**
Deep Models Combinations	DCL, DSS	0.888	0.875	0.875	0.590	0.876	0.890	0.894	**0.894**	0.886	0.893	**0.903**
	DCL, DSS, RFCN	0.888	0.837	0.838	0.814	0.844	0.894	0.895	**0.898**	0.874	0.891	**0.902**
	DCL, DSS, RFCN, MDF	0.888	0.845	0.845	0.816	0.852	0.887	0.892	**0.895**	0.874	0.886	**0.898**
	DCL, DSS, RFCN, MDF DHSNet	0.888	0.861	0.861	0.821	0.867	0.896	**0.903**	0.901	0.892	0.897	**0.905**
	DCL, DSS, RFCN, MDF MB+	0.888	0.842	0.842	0.799	0.851	0.890	0.891	**0.896**	0.877	0.886	**0.899**
	DCL, DSS, RFCN, MDF MB+,GP	0.888	0.838	0.838	0.793	0.846	0.884	0.884	**0.891**	0.872	0.881	**0.890**
	DCL, DSS, RFCN, MDF MB+,GP,MB	0.888	0.822	0.822	0.785	0.828	0.876	0.872	**0.889**	0.864	0.865	**0.884**
	DCL, DSS, RFCN, MDF MB+,BSCA,DRFI,TLLT	0.888	0.840	0.841	0.789	0.845	0.874	0.877	**0.889**	0.864	0.872	**0.886**

TABLE II: Mean F-measure of the average saliency maps (AVE), the resulted BN, M-estimator (M-est), and MCA saliency maps and the resulted AMS and AML saliency maps. The subscripts "B", "C" and "D" represent the boundary-based reference map, contour-based reference map and deep-network-based reference map respectively. The first column shows the combination strategy, and for every combination the highest F-measure of the candidate saliency models are displayed in the "Top" column. The **best result** for each combination is in bold with dark background color, the **second best** is in bold, and the third best is underlined. The candidate models include BSCA [19], CA [50], CEOS [68], COV [29], DRFI [58], FT [9], GBVS [5], GC [14], GP [45], HS [69], IS [31], IT [4], LR [32], MB [46], MB+ [46], MR [66], PCAS [30], RB [55], RC [14], SR [8], TLLT [23], UFO [58], DSS [70], DCL [71], RFCN [72], MDF [73], and DHSNet [44], of which deep models are underlined.

Dataset	Top	AVE	BN	M-est	MCA	AMS-B	AMS-C	AMS-D	AML-B	AML-C	AML-D
ECSSD	0.736	0.736	0.733	0.734	0.743	0.744	0.755	**0.790**	0.747	<u>0.770</u>	**0.825**
ASD	0.885	0.872	0.867	0.868	0.886	0.898	0.906	**0.917**	0.896	<u>0.912</u>	**0.928**
ImgSal	0.515	0.497	0.494	0.495	0.528	0.571	0.590	**0.636**	0.600	<u>0.626</u>	**0.690**
DUT-OMRON	0.546	0.560	0.556	0.556	0.571	0.602	0.616	**0.690**	0.621	<u>0.637</u>	**0.749**

TABLE III: Mean F-measure of the top saliency model, average saliency maps, BN, M-estimator (M-est), MCA model, and AM model with a combination of MB [46], BSCA [19], RC [14], GBVS [5], COV [29] and FT [9] models on four datasets including ECSSD [69], ASD [9], ImgSal [74] and DUT-OMRON [66]. The highest F-measure of the candidate saliency models are displayed in the "Top" column. The **best result** for each combination is in bold with dark background color, the **second best** is in bold, and the <u>third best</u> is underlined.

combination and so forth. Table II presents the evaluation results. Obviously, the AM model largely improves the F-measure of the top candidate saliency model with an average increase of 6.6%, 11.0%, 17.8%, 7.4%, 12.7%, and 20.9% for AMS-B, AMS-C, AMS-D, AML-B, AML-C, and AML-D correspondingly, while other online integration models such as AVE, BN, MCA and M-est decrease the F-measure by averagely 3.6%, 3.6%, 1.8% and 3.6% respectively. Apparently, the AM model greatly rectifies the error saliency information from the inferior candidate saliency models.

3. **Random combination.** From 2-model combination to 8-model combination, we randomly select candidate saliency models from the model pool and randomly evaluate five different combinations for each fixed number combination. The group "Random Models Combinations" in Table II shows one example of each fixed number combination with random selection strategy. Again, the AM model consistently outperforms each one of the combined saliency models and the MCA model. Figure 5 indicates the performance enhancement of the average maps, MCA model, AMS and AML model compared to the corresponding top models by averaging five random combinations for each fixed number combination. Apparently, the proposed model solidly improves the performance independent of the number of models being chosen for combination.

Also, we evaluate our AM model over four challenging datasets, ECSSD [69], ASD [9], ImgSal [74] and DUT-OMRON [66]. We use the combination of MB [46], BSCA [19], RC [14], GBVS [5], COV [29] and FT [9] as an example. Table III presents the F-measure of the top candidate saliency models, average saliency maps (AVE), BN, M-estimator, MCA results and our AM results over the four datasets. Our AM model largely improves the performance compared to the best candidate model on all the four datasets, and outperforms the average maps and MCA model all the time. Figure 6 (a)-(d) present the average MAE and F-measure of the candidate models being combined, the average saliency maps (AVE), results from BN, M-est and MCA model, and results from the AM model on the four datasets.

4. **Deep models combination.** In the experiment, we select deep models including DSS [70], DCL [71], RFCN [72], MDF [73], and DHSNet [44] for evaluation. In the last group "Deep Models Combinations" in Table II, we present different combinations involving deep models and traditional models. From the results, when deep models are involved as candidates, all the integration results of AM model that incorporate deep-network-based reference maps outperform the top candidate models. The first four rows are integration with all deep saliency models. The AMS-D and AML-D with deep external knowledge maps surpass the top saliency models averagely by 0.9% and 1.4% respectively. In the last four rows, the F-measure of the AVE, BN, MCA and M-est integration models drop sharply compared to the top models by averagely 5.3%, 5.2%, 9.7%, and 4.6% respectively, while AMS-D and AML-D averagely increase by 0.3% and 0.2% respectively.

In general, the AM model outperforms the existing integration models with all the four combination strategies. When the candidate saliency maps are from traditional models (*i.e.*, "Superior Models Combinations", "Inferior Models Combinations" and "Random Combinations"), AMS-B, AMS-C and AMS-D increase the F-measure of the top saliency models by averagely 3.1%, 5.0% and 10.0% respectively, while AML-B, AML-C and AML-D increase the F-measure of the top models by averagely 3.2%, 6.1% and 12.5% respectively. Thus, we conclude that the AM model solidly improves the performance regardless of the candidate models being combined.

Figure 7 shows some examples of the results of candidate saliency models, the average saliency maps, MCA model and the AM model on the four datasets.

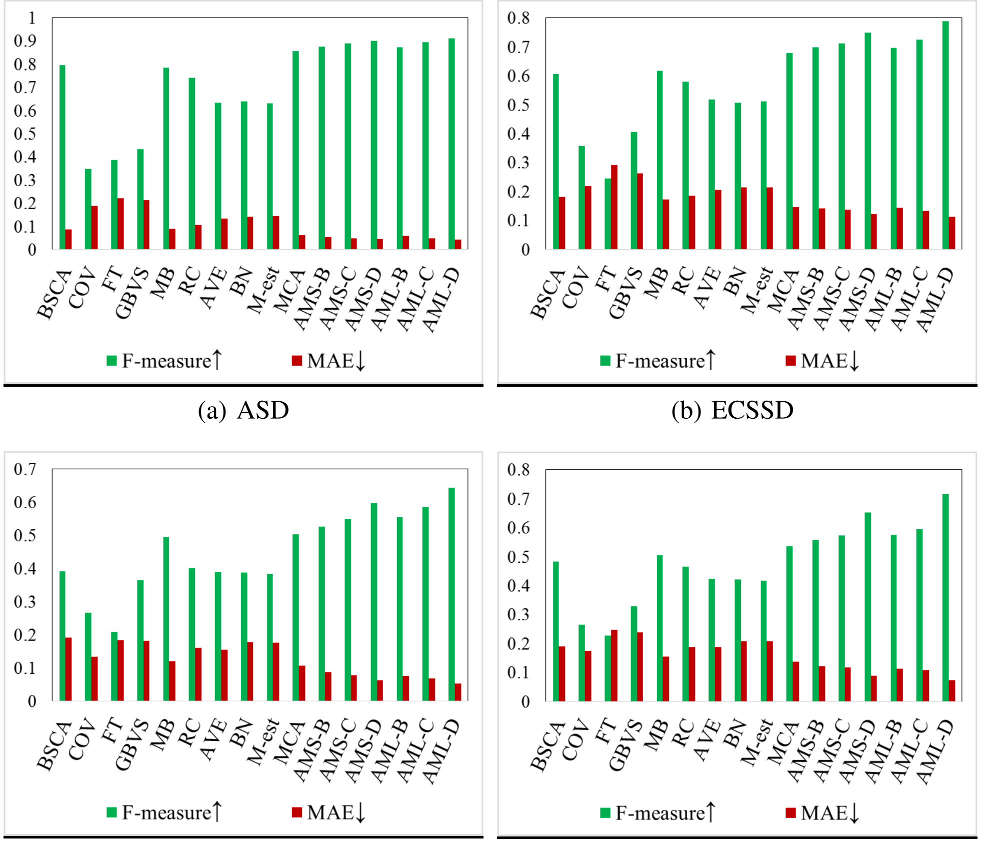
C. Rationality of the Reference Map and the Expertise

By evaluating the AM model with various combinations following four strategies, we conclude that the AM model substantially improves the performance regardless of the candidate models. In this section, we discuss the rationality of the reference map and the expertise respectively.

As mentioned in Section III, the reference map S_{Ref} is directly derived from $P(F)$ in Eq. 3, so that it should provide a natural scheme to introduce the information about salient object detection. In practice, the reference map is propagated from the consensus map of the external knowledge and the candidate saliency maps. Theoretically, the external knowledge can be any reasonable assumptions or currently existing models. In this paper, we address the inevitability of the reference map as one of the main components of the AM model. Thus, we firstly investigate the necessity of the reference map and then discuss the influences of different selections of the external knowledge.

We choose DRFI, GP, LR, MB+, TLLT and UFO as candidate saliency models and report the mean F-measure of the saliency integration results based on ECSSD dataset. As in Table V, the first row shows the performances when different external knowledge is incorporated to compute the reference

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2018.2856126, IEEE Transactions on Multimedia



(a) ASD

(b) ECSSD

(c) ImgSal

(d) DUT-OMRON

Fig. 6: (a)-(d) Average precision, recall, and F-measure of candidate saliency models, average saliency maps (AVE), BN, M-estimator (M-est), MCA and the AM model on the four datasets including ECSSD, ASD, ImgSal, and DUT-OMRON.

map, while the second row uses the external knowledge directly as an additional candidate saliency map. Obviously, the external knowledge being incorporated into the reference map results in better integration than it being integrated as another candidate map. Even when the external knowledge is a saliency model with inferior performance, incorporating the external knowledge for the reference map results in better integration results than taking the external knowledge as another candidate saliency model (first column of Table V). Thus, introducing the external knowledge receives performance enhancement in practice.

Further, from Table II, it can be perceived that when using the same expertise estimation method, the better the quality of the reference map, the higher performance the integration can be received. Apparently, the incorporation of the reference map is critical to the performance enhancement.

As mentioned in Section III, during the CA updating process, the reference map is updated by averaging the candidate saliency maps. Thus, we conduct a small scale experiment to evaluate whether the reference map should be updated at every iteration. We test with three combinations on ECSSD dataset: 1) DHSNet, GP, IT, 2) GP, LR, PCAS, and 3) GC, GP, PCAS. By updating the reference map in each generation, the mean F-measure of the above three combinations are higher than keeping the reference map unchanged by averagely 2.01%, 0.23% and 0.06% for AMS-B, AMS-C, AMS-D and 1.30%, 0.33%, 0.07% for AML-B, AML-C, AML-D respectively. Thus, we finally decide to update the reference map by considering both the theoretical inference and the practice.

According to Table II, AML performs similarly well or slightly better than AMS in general when the same external knowledge is introduced. However, there are some exceptions. For instance, the AMS-B outperforms the AML-B by averagely 2.2% in “Deep Models Combinations”. The reason behind this is that different from the latent-variable-based expertise, the statistics-based expertise borrows information from the reference map as the approximated ‘Ground Truth’ as in Section IV-A. Thus, the accuracy of the statistics-based expertise is influenced by the quality of the reference map. In “Deep Models Combinations”, the high performances of deep candidates largely enhance the quality of the reference map, such that the AMS-B outperforms the AML-B significantly.

To quantitatively investigate the contributions of the reference map and the expertise, we also experiment each unit of the AM model on ECSSD dataset with a combination of candidate models including DRFI, GP, LR, MB+, TLLT and UFO, as shown in Table IV. The basic framework is an integration of candidate saliency models with equal expertise but without the reference map based on CA. Then we add different components to the basic framework, where the letter

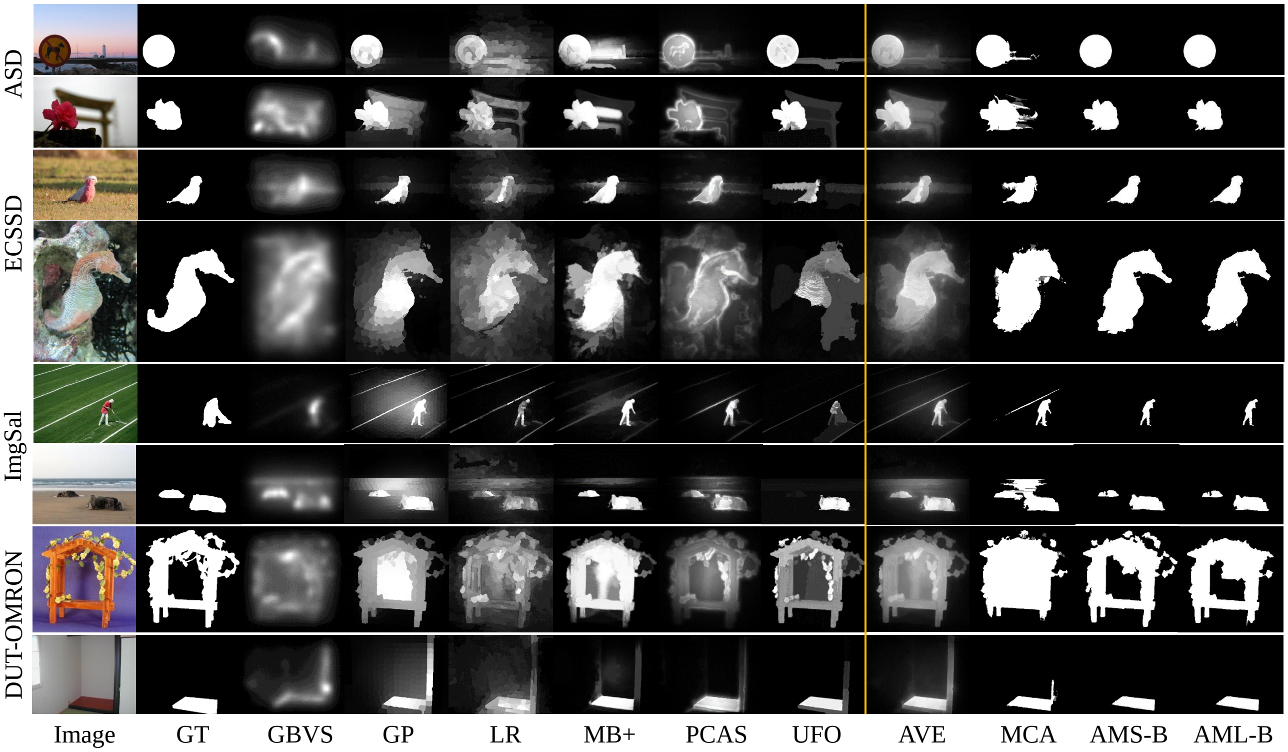


Fig. 7: Examples of the results of combined saliency models, average saliency maps (AVE), BN, M-estimator (M-est), MCA, AMS and AML. The images with ground truth (GT) are sequentially from ECSSD, ASD, ImgSal, and DUT-OMRON datasets.

"L" refers to the latent-variable-based expertise, "S" means the statistics-based expertise, and "Ref-B" uses the boundary-prior-based external knowledge. From Table IV, it is obvious that the introduction of the reference map significantly enhances the performance of the basic CA framework. The involvement of the statistics-based expertise and the latent-variable-based expertise also improves the performance of the basic framework respectively. Finally, the statistics-based integration and the latent-variable-based integration with the reference map result in similarly better performances than only incorporating single unit in CA. Thus, the incorporation of both the reference map and the expertise of saliency models results in the best performance.

The proposed AM model synchronizes the p-th candidate by using its continuous map and the other candidates as binary maps during the CA process. Such updating form keeps the actual saliency intensity of the current candidate map and eliminates diverse semantic meanings and magnitudes from the other saliency models, as mentioned in Section IV. In Table VI, we present the mean F-measure of the integration results by using 1) only continuous maps for integration, 2) only binary maps for integration, and 3) both continuous and binary maps for integration. It can be perceived that using both the continuous maps and binary maps produces the best integration results. Although the resulted maps by using both continuous and binary maps are only slightly better than or equal to the results by using only binary maps, it does not increase the computation complexity and even reduces the numbers of thresholding process once per iteration.

Ref-B	×	×	×	√	√	√
Expertise	×	L	S	×	L	S
F-Measure	0.757	0.767	0.762	0.769	0.777	0.784

TABLE IV: Mean F-measure of saliency integration framework incorporating various combination of the proposed components by AM model. The letter "L" refers to the latent-variable-based expertise and "S" means the statistics-based expertise. "Ref-B" means that we test the boundary-based reference map. "$\surd$" and "$\times$" indicate whether the component is incorporated into CA or not.

External Knowledge	FT	Bound	CCM	DHSNet
Reference	0.766	0.789	0.777	0.837
Candidate	0.744	0.750	0.757	0.774

TABLE V: Mean F-measure of saliency integration results of the AML model by involving different external knowledge. The first row ("Reference") and the second row ("Candidate") refer to the F-measure by incorporating the external knowledge as the reference map and as one of the candidate maps to be aggregated, respectively.

D. Discussion of Convergence

As mentioned in Section II-B, the synchronizing updating rule of the cellular automaton is designed to converge the evolved cells to a stable state after several generations. We compute the absolute difference of the S^t_{Ref} and S^{t-1}_{Ref} in Eq. 5 at each generation, and plot the average absolute

	AML			AMS		
	B	C	D	B	C	D
Binary	0.777	0.787	0.833	0.781	0.788	0.822
Continuous	0.759	0.74	0.762	0.603	0.511	0.509
Both	0.777	0.789	0.837	0.781	0.788	0.823

TABLE VI: Mean F-measure of saliency integration by using AM model as framework with only continuous saliency maps, only binary maps, and both continuous and binary maps for integration. The candidate saliency models are DRFI, GP, LR, MB+, TLLT and UFO on ECSSD dataset.

difference between S^t_{Ref} and S^{t-1}_{Ref} of all the superpixels on one image, with a combination of MB [46], BSCA [19], RC [14], GBVS [5], COV [29] and FT [9] on ECSSD dataset. The result is illustrated on Figure 5 (c). As is shown, the designed updating rule for cellular automaton can make the S^t_{Ref} rapidly converge within five iterations.

E. Running Time

We implement our method in MATLAB R2014b using a Windows desktop with an i5-3570 CPU at 3.40GHz. The running time of AMS-B on ECSSD dataset ranges from 1.28s (2-model-combination) to 1.32s (8-model-combination) per image, while AML-B ranges from 1.38s (2-model-combination) to 2.06s (8-model-combination) in average, without code optimization. The AML and AMS show comparable performances, but AML takes longer time in running the EM algorithm.

VI. Conclusion

This paper presents an arbitrator model (AM) as an efficient online saliency integration model to release the burden of model-training from offline models. On one hand, the AM model introduces the reference map to overcome the misleading of inferior saliency models by exploring the consensus of the multiple saliency maps and the external knowledge. On the other hand, it rationally learns the expertise of saliency models without any knowledge of the ground truth labels in an online manner. We evaluate the AM with a pool of twenty-seven models under various combinations. The experimental results show that it substantially improves the performance, regardless of the choices of candidate approaches.

We also hold discussions about the two proposed online methods in estimating the expertise of saliency models, namely the statistics-based expertise and the latent-variable-based expertise. It can be easily observed that the statistics-based expertise is more accurate than the latent-variable-based one, if without the reference map. With the incorporation of the reference map, the two proposed expertise estimation methods perform similarly well. Nevertheless, the computational cost of the latent-variable-based method is higher than that of the statistics-based method, especially when the number of candidate models increases. Therefore, the statistics-based expertise is more efficient than the latent-variable-based expertise when large numbers of saliency models are integrated.

The AM model proposes a new integration framework that incorporates the reference map and the candidate saliency models of varying expertise. Although the framework is derived from Bayesian inference, the computation of each component can be further investigated. Firstly, the design of each component can be explored for further gains in performance. This paper incorporates the consensus of multiple saliency models and the external knowledge to approximate the reference map. To further improve the quality of the reference map, the external knowledge can be introduced in more complex ways, *i.e.*, using multiple external knowledge rather than one to increase the validity of the reference map. Also, the multiple saliency models could contribute to the reference map in other forms rather than the majority voting, such as adopting mean field approximation [77] to estimate the reference map. Secondly, the expertise can also be evaluated in different forms. In this work, we suggest to use the latent-variable-based method if the quality of the reference map is poor and the statistics-based method if the reference map is powerful. However, the latent-variable-based approach and the statistics-based approach can also be combined to stabilize the accuracy of expertise estimation if the quality of the reference map is uncertain and the computational cost is out of consideration.

Currently, at the updating stage of the cellular automaton in AM model, the state of a cell is only affected by the superpixels at the same location of all the saliency maps. In future, we will explore the influences of the adjacent superpixels of a cell. Also, the reference map and the expertise estimation of the AM framework can also be applied to co-saliency detection [78], [79]. Moreover, we could further apply the framework to other tasks such as anomaly detection [80].

Acknowledgment

The authors are grateful to the Academy of Finland, Tekes Fidipro Program (Grant No. 1849/31/2015), Business Finland project (Grant No.3116/31/2017), Infotech, and Natural Science Foundation of China (Grant No. 61772419/61572205/61601362). The authors also wish to acknowledge CSC-IT Center for Science, Finland, for generous computational resources, and the supports of NVIDIA Corporation with the donation of the GPUs used for this research. Besides, we appreciate the comments and codes from Prof. Huchuan Lu and Mr. Mengyang Feng.

References

[1] E. Rahtu, J. Kannala, M. Salo, and J. Heikkilä, "Segmenting salient objects from images and videos," in *Proc. ECCV*, 2010, pp. 366–379.

[2] C. Guo and L. Zhang, "A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression," *TIP*, vol. 19, no. 1, pp. 185–198, 2010.

[3] X.-S. Hua, T. Mei, and S. Li, "When multimedia advertising meets the new internet era," in *Workshop on Multimedia Signal Processing*. IEEE, 2008, pp. 1–5.

[4] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *TPAMI*, vol. 20, no. 11, pp. 1254–1259, 1998.

[5] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *NIPS*, 2006, pp. 545–552.

[6] D. Walther and C. Koch, "Modeling attention to salient proto-objects," *Neural Networks*, vol. 19, no. 9, pp. 1395–1407, 2006.

[7] N. Bruce and J. Tsotsos, "Saliency based on information maximization," in *NIPS*, 2005, pp. 155–162.

[8] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *Proc. CVPR*. IEEE, 2007, pp. 1–8.

[9] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Proc. CVPR*. IEEE, 2009, pp. 1597–1604.

[10] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *Proc. ICCV*. IEEE, 2009, pp. 2106–2113.

[11] M. Jiang, J. Xu, and Q. Zhao, "Saliency in crowd," in *Proc. ECCV*. Springer, 2014, pp. 17–32.

[12] O. L. Meur, P. L. Callet, D. Barba, and D. Thoreau, "A coherent computational approach to model bottom-up visual attention," *TPAMI*, vol. 28, no. 5, pp. 802–817, 2006.

[13] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, "Sun: A bayesian framework for saliency using natural statistics," *Journal of Vision*, vol. 8, no. 7, p. 32, 2008.

[14] M. Cheng, N. J. Mitra, X. Huang, P. H. Torr, and S. Hu, "Global contrast based salient region detection," *TPAMI*, vol. 37, no. 3, pp. 569–582, 2015.

[15] S. Lu, V. Mahadevan, and N. Vasconcelos, "Learning optimal seeds for diffusion-based salient object detection," in *Proc. CVPR*. IEEE, 2014, pp. 2790–2797.

[16] W. Zhu, S. Liang, Y. Wei, and J. Sun, "Saliency optimization from robust background detection," in *Proc. CVPR*. IEEE, 2014, pp. 2814–2821.

[17] T. Shi, M. Liang, and X. Hu, "A reverse hierarchy model for predicting eye fixations," in *Proc. CVPR*. IEEE, 2014, pp. 2822–2829.

[18] R. Liu, J. Cao, Z. Lin, and S. Shan, "Adaptive partial differential equation learning for visual saliency detection," in *Proc. CVPR*. IEEE, 2014, pp. 3866–3873.

[19] Y. Qin, H. Lu, Y. Xu, and H. Wang, "Saliency detection via cellular automata," in *Proc. CVPR*. IEEE, 2015, pp. 110–119.

[20] N. Liu, J. Han, D. Zhang, S. Wen, and T. Liu, "Predicting eye fixations using convolutional neural networks," in *Proc. CVPR*. IEEE, 2015, pp. 362–370.

[21] R. Zhao, W. Ouyang, H. Li, and X. Wang, "Saliency detection by multi-context deep learning," in *Proc. CVPR*. IEEE, 2015, pp. 1265–1274.

[22] N. Tong, H. Lu, X. Ruan, and M.-H. Yang, "Salient object detection via bootstrap learning," in *Proc. CVPR*. IEEE, 2015, pp. 1884–1892.

[23] C. Gong, D. Tao, W. Liu, S. J. Maybank, M. Fang, K. Fu, and J. Yang, "Saliency propagation from simple to difficult," in *Proc. CVPR*. IEEE, 2015, pp. 2531–2539.

[24] D. Zhang, J. Han, C. Li, and J. Wang, "Co-saliency detection via looking deep and wide," in *Proc. CVPR*. IEEE, 2015, pp. 2994–3002.

[25] M. Feng, A. Borji, and H. Lu, "Fixation prediction with a combined model of bottom-up saliency and vanishing point," in *WACV*. IEEE, 2016, pp. 1–7.

[26] Y. Xu, X. Hong, Q. He, G. Zhao, and M. Pietikäinen, "A task-driven eye tracking dataset for visual attention analysis," in *International Conference on Advanced Concepts for Intelligent Vision Systems*. Springer, 2015, pp. 637–648.

[27] L. Wang, H. Lu, X. Ruan, and M.-H. Yang, "Deep networks for saliency detection via local estimation and global search," in *Proc. CVPR*. IEEE, 2015, pp. 3183–3192.

[28] C. Kanan, M. H. Tong, L. Zhang, and G. W. Cottrell, "Sun: Top-down saliency using natural statistics," *Visual Cognition*, vol. 17, no. 6-7, pp. 979–1003, 2009.

[29] E. Erdem and A. Erdem, "Visual saliency estimation by nonlinearly integrating features using region covariances," *Journal of Vision*, vol. 13, no. 4, p. 11, 2013.

[30] R. Margolin, A. Tal, and L. Zelnik-Manor, "What makes a patch distinct?" in *Proc. CVPR*. IEEE, 2013, pp. 1139–1146.

[31] X. Hou, J. Harel, and C. Koch, "Image signature: Highlighting sparse salient regions," *TPAMI*, vol. 34, no. 1, pp. 194–201, 2012.

[32] X. Shen and Y. Wu, "A unified approach to salient object detection via low rank matrix recovery," in *Proc. CVPR*. IEEE, 2012, pp. 853–860.

[33] Y. Xu, X. Hong, X. Liu, and G. Zhao, "Saliency detection via bi-directional propagation," *Journal of Visual Communication and Image Representation*, vol. 53, pp. 113–121, 2018.

[34] C. Shen and Q. Zhao, "Webpage saliency," in *Proc. ECCV*, 2014, pp. 33–46.

[35] N. D. Bruce, C. Catton, and S. Janjic, "A deeper look at saliency: Feature contrast, semantics, and beyond," in *Proc. CVPR*. IEEE, 2016, pp. 516–524.

[36] J. Pan, K. McGuinness, E. Sayrol, N. O'Connor, and X. Giro-i Nieto, "Shallow and deep convolutional networks for saliency prediction," in *Proc. CVPR*. IEEE, 2016.

[37] O. Le Meur and Z. Liu, "Saliency aggregation: Does unity make strength?" in *Proc. ACCV*. Springer, 2014, pp. 18–32.

[38] R. Song, Y. Liu, R. R. Martin, and P. L. Rosin, "Saliency-guided integration of multiple scans," in *Proc. CVPR*. IEEE, 2012, pp. 1474–1481.

[39] L. Mai, Y. Niu, and F. Liu, "Saliency aggregation: a data-driven approach," in *Proc. CVPR*. IEEE, 2013, pp. 1131–1138.

[40] J. Wang, A. Borji, C.-C. J. Kuo, and L. Itti, "Learning a combined model of visual saliency for fixation prediction," *TIP*, vol. 25, no. 4, pp. 1566–1579, 2016.

[41] A. Borji, D. N. Sihite, and L. Itti, "Salient object detection: A benchmark," in *Proc. ECCV*. Springer, 2012, pp. 414–429.

[42] L. Mai and F. Liu, "Comparing salient object detection results without ground truth," in *Proc. ECCV*. Springer, 2014, pp. 76–91.

[43] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, "Salient object detection: A benchmark," *TIP*, vol. 24, no. 12, pp. 5706–5722, 2015.

[44] N. Liu and J. Han, "Dhsnet: Deep hierarchical saliency network for salient object detection," in *Proc. CVPR*, 2016, pp. 678–686.

[45] P. Jiang, N. Vasconcelos, and J. Peng, "Generic promotion of diffusion-based salient object detection," in *Proc. ICCV*. IEEE, 2015, pp. 217–225.

[46] J. Zhang, S. Sclaroff, Z. Lin, X. Shen, B. Price, and R. Mech, "Minimum barrier salient object detection at 80 fps," in *Proc. ICCV*. IEEE, 2015, pp. 1404–1412.

[47] R. Trichet and N. E. O'Connor, "A flexible ensemble-svm for computer vision tasks," in *Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, 2016.

[48] D. Zhang, J. Han, and Y. Zhang, "Supervision by fusion: Towards unsupervised learning of deep salient object detector," in *Proc. ICCV*. IEEE, 2017, pp. 4048–4056.

[49] R. Quan, J. Han, D. Zhang, F. Nie, X. Qian, and X. Li, "Unsupervised salient object detection via inferring from imperfect saliency models," *TMM*, vol. 1, p. 1, 2017.

[50] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," *TPAMI*, vol. 34, no. 10, pp. 1915–1926, 2012.

[51] N. Otsu, "A threshold selection method from gray-level histograms," *Automatica*, vol. 11, no. 285-296, pp. 23–27, 1975.

[52] J. Von Neumann, "The general and logical theory of automata," *Cerebral Mechanisms in Behavior*, pp. 1–41, 1951.

[53] S. Wolfram, "Statistical mechanics of cellular automata," *Reviews of Modern Physics*, vol. 55, no. 3, p. 601, 1983.

[54] Y. Qin, M. Feng, H. Lu, and G. W. Conttrell, "Hierarchical cellular automata for visual saliency," *IJCV*, pp. 1–20, 2018.

[55] Y. Wei, F. Wen, W. Zhu, and J. Sun, "Geodesic saliency using background priors," in *Proc. ECCV*. Springer, 2012, pp. 29–42.

[56] X. Li, H. Lu, L. Zhang, X. Ruan, and M.-H. Yang, "Saliency detection via dense and sparse reconstruction," in *Proc. ICCV*. IEEE, 2013, pp. 2976–2983.

[57] Q. Wang, W. Zheng, and R. Piramuthu, "Grab: Visual saliency via novel graph model and background priors," in *Proc. CVPR*. IEEE, 2016, pp. 535–543.

[58] J. Wang, H. Jiang, Z. Yuan, C. Ming-Ming, X. Hu, and N. Zheng, "Salient object detection: A discriminative regional feature integration approach," *IJCV*, vol. 123, no. 2, pp. 251–268, 2017.

[59] J. Han, D. Zhang, X. Hu, L. Guo, J. Ren, and F. Wu, "Background prior-based salient object detection via deep reconstruction residual," *TCSVT*, vol. 25, no. 8, pp. 1309–1321, 2015.

[60] D. Zhang, J. Han, L. Jiang, S. Ye, and X. Chang, "Revealing event saliency in unconstrained video collection," *TIP*, vol. 26, no. 4, pp. 1746–1758, 2017.

[61] Q. Liu, X. Hong, B. Zou, J. Chen, Z. Chen, and G. Zhao, "Hierarchical contour closure based holistic salient object detection," *TIP*, 2017.

[62] J. A. Hartigan and M. A. Wong, "Algorithm as 136: A k-means clustering algorithm," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 28, no. 1, pp. 100–108, 1979.

[63] V. Gopalakrishnan, Y. Hu, and D. Rajan, "Random walks on graphs for salient object detection in images," *TIP*, vol. 19, no. 12, pp. 3232–3242, 2010.

[64] B. Jiang, L. Zhang, H. Lu, C. Yang, and M.-H. Yang, "Saliency detection via absorbing markov chain," in *Proc. ICCV*. IEEE, 2013, pp. 1665–1672.

[65] Z. Ren, Y. Hu, L.-T. Chia, and D. Rajan, "Improved saliency detection based on superpixel clustering and saliency propagation," in *Proc. Multimedia*. ACM, 2010, pp. 1099–1102.

[66] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *Proc. CVPR*, 2013, pp. 3166–3173.

[67] J. Whitehill, T.-f. Wu, J. Bergsma, J. R. Movellan, and P. L. Ruvolo, "Whose vote should count more: Optimal integration of labels from labelers of unknown expertise," in *NIPS*, 2009, pp. 2035–2043.

[68] R. Mairon and O. Ben-Shahar, "A closer look at context: From coxels to the contextual emergence of object saliency," in *Proc. ECCV*. Springer, 2014, pp. 708–724.

[69] J. Shi, Q. Yan, L. Xu, and J. Jia, "Hierarchical image saliency detection on extended cssd," *TPAMI*, vol. 38, no. 4, pp. 717–729, 2016.

[70] Q. Hou, M.-M. Cheng, X.-W. Hu, A. Borji, Z. Tu, and P. Torr, "Deeply supervised salient object detection with short connections," *TPAMI*, 2018.

[71] G. Li and Y. Yu, "Deep contrast learning for salient object detection," in *Proc. CVPR*. IEEE, 2016.

[72] J. Dai, Y. Li, K. He, and J. Sun, "R-fcn: Object detection via region-based fully convolutional networks," in *NIPS*, 2016, pp. 379–387.

[73] G. Li and Y. Yu, "Visual saliency based on multiscale deep features," in *Proc. CVPR*. IEEE, 2015, pp. 5455–5463.

[74] J. Li, M. D. Levine, X. An, X. Xu, and H. He, "Visual saliency based on scale-space analysis in the frequency domain," *TPAMI*, vol. 35, no. 4, pp. 996–1010, 2013.

[75] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.-Y. Shum, "Learning to detect a salient object," *TPAMI*, vol. 33, no. 2, pp. 353–367, 2011.

[76] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *TPAMI*, vol. 34, no. 11, pp. 2274–2282, 2012.

[77] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," in *Proc. NIPS*, 2011, pp. 109–117.

[78] D. Zhang, D. Meng, and J. Han, "Co-saliency detection via a self-paced multiple-instance learning framework," *TPAMI*, vol. 39, no. 5, pp. 865–878, 2017.

[79] X. Yao, J. Han, D. Zhang, and F. Nie, "Revisiting co-saliency detection: A novel approach based on two-stage multi-view spectral rotation co-clustering," *TIP*, vol. 26, no. 7, pp. 3196–3209, 2017.

[80] K. J. Ponti, Moacir and, M. Riva, d. T. deCampos, and C. Zor, "A decision cognizant kullbackleibler divergence," *Pattern Recognition*, vol. 61, pp. 470–478, 2017.

Yingyue Xu is a Ph.D. candidate with the Center for Machine Vision and Signal Analysis, University of Oulu, Finland. She received her B.Eng degree in Liaoning University, China in 2011 and then worked with Neusoft Corporation on car navigation until 2013. She received her M.Sc degree in Computer Science and Engineering focusing on Computer Vision and Image Processing in 2015. Her research interests include visual attention analysis, human visual behavior analysis, and saliency detection.

Xiaopeng Hong (M'13) received his B.Eng and Ph.D. degrees from Harbin Institute of Technology, Harbin, China, in 2004 and 2010 respectively. He is a Docent with the Center for Machine Vision and Signal Analysis, University of Oulu, Finland, where he has been a senior researcher since 2011. He has published over 30 articles in journals and conferences such as IEEE T-PAMI and IEEE CVPR. He has organized two workshops and served as a reviewer for 20 journals. His research interests include multi-modal learning, affective computing, medical examination, and human-computer interaction. His research has been reported by media such as *MIT Technology Review* and *Daily Mail*.

Fatih Porikli (M'96,SM'04,F'14) is an IEEE Fellow and a Professor in the Research School of Engineering, Australian National University (ANU). He is also acting as the Chief Scientist at Huawei, Santa Clara. He has received his Ph.D. from New York University in 2002. Previously he served Distinguished Research Scientist at Mitsubishi Electric Research Laboratories. His research interests include computer vision, pattern recognition, manifold learning, image enhancement, robust and sparse optimization and online learning with commercial applications in video surveillance, car navigation, intelligent transportation, satellite, and medical systems.

Xin Liu (M'16) is a Ph.D. candidate with the Center for Machine Vision and Signal Analysis, University of Oulu, Finland. He received the B.Sc. and M.Sc. degrees in Computer Science in 2003 and 2007 respectively. His research interests include human behavior analysis, 3D computer vision, image restoration, and object detection.

Jie Chen (M'05) received his M.S. and Ph.D. degrees from Harbin Institute of Technology, China, in 2002 and 2007, respectively. Since 2007, he has been a senior researcher in the Machine Vision Group at the University of Oulu, Finland. In 2012 and 2015, he visited the Computer Vision Laboratory at University of Maryland and School of Electrical and Computer Engineering at Duke University respectively. Dr. Chen was a co-chair of International Workshops at ACCV, CVPR, and ICCV. He was a guest editor of special issues for IEEE TPAMI, IJCV and Neurocomputing. His research interests include pattern recognition, computer vision, machine learning, dynamic texture, deep learning, and medical image analysis. He is an Associate Editor of The Visual Computer.

Guoying Zhao (SM'12) is a Professor with the Center for Machine Vision and Signal Analysis, University of Oulu, Finland, where she has been a senior researcher since 2005 and an Associate Professor since 2014. She received the Ph.D. degree in computer science from the Chinese Academy of Sciences, China, in 2005. In 2011, she was selected to the highly competitive Academy Research Fellow position. She was Nokia visiting professor in 2016. She has authored or co-authored more than 180 papers in journals and conferences. Her papers have currently over 7700 citations in Google Scholar (h-index 37). She is co-publicity chair for FG2018, has served as area chairs for several conferences and is associate editor for Pattern Recognition, IEEE T-CSVT, and IVC Journals. She has lectured tutorials at ICPR 2006, ICCV 2009, SCIA 2013 and FG 2018, authored/edited three books and eight special issues in journals. Dr. Zhao was a Co-Chair of many International Workshops at ECCV, ICCV, CVPR, ACCV and BMVC. Her current research interests include image and video descriptors, facial-expression and micro-expression recognition, gait analysis, dynamic-texture recognition, human motion analysis, and person identification. Her research has been reported by Finnish TV programs, newspapers and MIT Technology Review.