IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 18, 2020, accepted June 4, 2020, date of publication June 9, 2020, date of current version June 26, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3001041

Inaccurate Supervised Saliency Detection Based on Iterative Feedback Framework

YU PANG[1], YUNHE WU[2], CHENGDONG WU[1], XIAOSHENG YU[1], AND YUAN GAO[2]
[1]Faculty of Robot Science and Engineering, Northeastern University, Shenyang 110169, China
[2]College of Information Science and Engineering, Northeastern University, Shenyang 110819, China

Corresponding authors: Chengdong Wu (wuchengdong@ise.neu.edu.cn) and Xiaosheng Yu (yuxiaoshengneu@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61701101, Grant 61973093, Grant U1713216, Grant 61901098, and Grant 61971118, in part by the Fundamental Research Fund for the Central Universities of China under Grant N2026005, Grant N181602014, Grant N2026004, Grant N2026006, Grant N2026001, and Grant N2011001, and in part by the Project for the Science and Technology Major Special Plan of Liaoning under Grant 2019JH1/10100005.

ABSTRACT Machine learning based bottom-up saliency detection (MLBU) methods are very popular recently. These MLBU methods firstly use prior knowledge to select some regions from the given image as training samples and label them. Based on training set, a saliency classifier is learned to classify salient object and background by applying machine learning algorithms in the given image. Nevertheless, training labels obtained by prior knowledge are not always accurate in some complex scenes, inaccurate training set is hard to make subsequent learning process succeed. To solve this problem, we propose an inaccurate supervised learning (ISL) based saliency detection framework, which assumes that training labels obtained by prior knowledge might be inaccurate and constructs three checking rules to remove mislabeled samples for more accurate training set construction. The refined training set is used to learn a saliency classifier which can better predict each image region. To obtain more accurate saliency inference, the proposed ISL process is introduced into a novel iterative feedback (IF) framework to generate better saliency result. Finally, we use smoothness operator to further smooth saliency result for performance improvement. Experimental results on three benchmark datasets demonstrate adequately the superiority of the proposed method.

INDEX TERMS Saliency detection, prior knowledge, inaccurate supervised learning, iterative feedback classification, smoothness optimization.

I. INTRODUCTION

Saliency detection has become an important topic in computer vision and image processing tasks. Its goal is to identify the most interesting regions that attract human eye attention in a natural scene. Thus, saliency detection is widely applied to numerous computer vision and image processing applications, such as image classification [1], scene recognition [2], image segmentation [3] and so on. More and more researchers focus on saliency detection field. Generally speaking, state-of-the-art methods are divided into two strategies, i.e., top-down (TD) methods and bottom-up (BU) methods.

Top-down methods are usually driven by specific tasks, they need to learn saliency model from numerous training images with the ground truth. Deep neural network (DNN) based methods [7]–[10] are the most popular top-down methods, which can better exploit the high-level semantic information of image due to hierarchical architecture. Thus, DNN based methods have achieved the promising performances recently. However, for DNN based methods, collecting training images with manual annotation is a time-consuming work. As a result, using DNN for saliency detection, although effective, is relatively less economical than bottom-up approaches.

In contrast, bottom-up methods are faster and simpler than top-down methods, because training images with manual annotation are not needed. These methods usually exploit saliency cues by utilizing various prior knowledge, e.g., background prior, center prior, contrast prior, spatial prior and so on. However, prior knowledge only can provide a coarse but imprecise indicator in most scenes. Based on this observation, machine learning algorithms are widely applied to bottom-up methods (i.e., MLBU methods), which usually

The associate editor coordinating the review of this manuscript and approving it for publication was Huiling Chen.

contain two stages: training set construction, saliency classifier construction and prediction. For an input image, the first stage is to use prior knowledge to select some regions from input image as training samples and then give them training labels (positive label for ''foreground'' and negative label for ''background''), it's noticed that various MBLU methods may use different prior knowledge. In the second stage, the selected training set is used to learn a saliency classifier based on different machine learning algorithms, such as bootstrap learning [4], multi-instance learning [5], dictionary learning [6] and so on. The learned saliency classifier can predict each image region to be foreground or background. As a result, a binary saliency map is obtained by MLBU strategy.

Different from top-down (TD) methods, the labeling task of training samples in MLBU methods is determined by prior knowledge, which is more time-saving than top-down ones. Nevertheless, a vital drawback is that training labels obtained by prior knowledge might be inaccurate, especially in some complex scenes. Because various prior knowledge only provide coarse but imprecise indicators for subsequent learning process. It means that a terrible training set easily lead to bad saliency result. In other words, strongly supervised information is hard to obtain in MLBU methods.

Inaccurate supervised learning, which is one of the most important weakly supervised learning frameworks, assumes that a subset of training set is mislabeled and aims to construct different rules to check and correct inaccurate samples. This setting is widely applied to various learning tasks, such as image classification, face recognition and so on. We find that inaccurate supervised learning framework also contributes to MLBU methods, because strongly supervised information is not always obtained in MLBU strategy.

Based on above observation, we attempt to solve this problem about existing MLBU methods by proposing an iterative feedback based inaccurate supervised learning (IFISL) framework. (1)For one thing, we exploit saliency cues based on an inaccurate supervised learning framework (ISL). For input image, we firstly construct a coarse training set by integrating three well-known prior knowledge, i.e., background prior, global contrast prior and objectness prior. Secondly, the ISL assumes that a subset of training labels might be inaccurate due to the limitation of prior knowledge. We are surprised to find that this setting is of great importance but studied rarely in bottom-up methods. To achieve this goal, we propose three checking rules (i.e., local consistency rule, feature contrast rule and spatial distribution rule) to check the label reliabilities of all training samples and remove mislabeled samples from training set. Thirdly, more refined training set is used to learn a saliency classifier to classify each region to be foreground/background in the given image. (2) For the other thing, in order to obtain better saliency result, we introduce the proposed ISL process into a novel iterative feedback framework (IF), in which each iteration is a ISL process with feedback mechanism. In our IF framework, saliency result is updated constantly to the optimized stable state, which is associated with an accurate saliency map. Finally, we use smoothness operator to further smooth saliency map for performance improvement. To better emphasize the difference between the proposed IFISL and MLBU framework, we show respectively their frameworks in Fig.1.

In summary, the contributions of our work are listed as follows:

- We are the first to introduce inaccurate supervised learning (ISL) into MLBU framework. The ISL is able to solve the problem that training labels might be inaccurate in MLBU by constructing three checking rules, and generate better saliency result than conventional MLBU methods, especially in some complex scenes.
- We introduce the proposed ISL framework into a novel iterative feedback (IF) framework, the introduction of feedback regulation mechanism can further improve the quality of saliency result.

II. RELATED WORK

Deep neural network (DNN) based methods are the most popular top-down methods, which have achieved top performances recently. e.g., Wang *et al.* [7] construct two deep neural networks to exploit saliency cues based on global and local perspectives, they are respectively global search network and local estimation network. He *et al.* [8] propose a novel network named Super-CNN to generate superpixel-level saliency map. Qin *et al.* [9] attempt to use pre-trained network to extract deep features which can exploit the high-level semantic information of image. Zeng *et al.* [10] construct an iterative random walk model to take both advantages of low-level features and high-level features. However, the annotation work is very tedious for top-down methods, and training sets with accurate annotations remain scarce and expensive. To solve this problem, weak supervision information is applied to top-down methods containing DNN based methods, which aim to train a top-down model when we do not have sufficient labeled data or only have limited labels. e.g., Wang *et al.* [33] provide a new paradigm for learning saliency detectors with weak supervision, which only requires less annotation efforts. Qian *et al.* [34] propose a novel feature matching network based on weak supervision framework to explore the natural relationship between language and image, which provides an important saliency prior for detection. Zeng *et al.* [40] attempt to train a classification network based on multiple weak supervision sources. Hsu *et al.* [41] learn a classifier-driven map generator under weak supervision framework. Besides, Tang *et al.* [35] propose a deep saliency quality assessment network which can better evaluate the quality of saliency map.

In contrast, some bottom-up methods detect salient object based on various prior knowledge. As one of the most popular priors in saliency detection, background prior [11]–[13], [28]–[30] firstly defines image boundary regions to be background seeds, and then each region's saliency value is its feature contrast with the background seeds. e.g., Zhang *et al.* [11] compute each region's manifold ranking score with image boundary regions to represent its

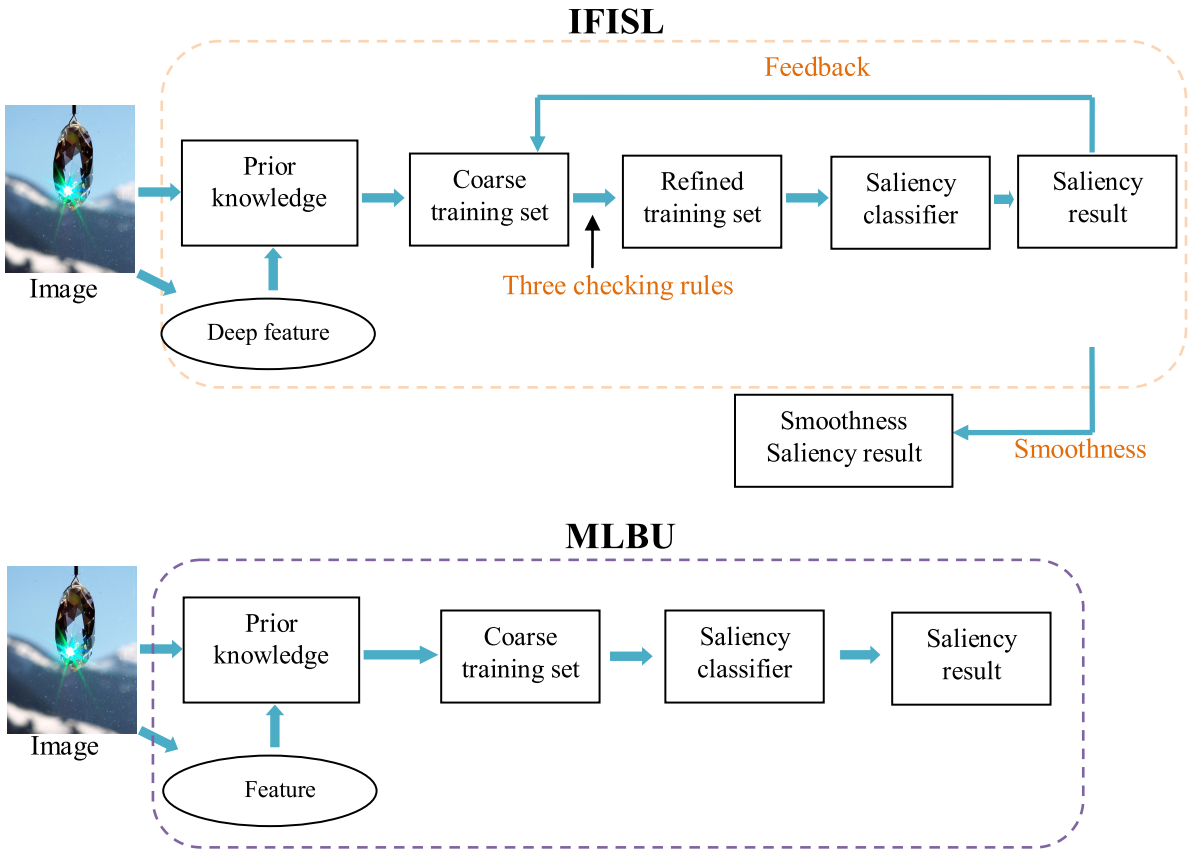

FIGURE 1. The flows of conventional MLBU methods and the proposed IFISL. There are two contributions in the proposed IFISL:(1)Three checking rules are proposed to check label reliabilities to generate refined training set (2) An iterative feedback mechanism is introduced into learning process.

saliency value. Li *et al.* [12] construct a label propagation mechanism which propagates saliency value from image boundary regions to other regions. Contrast prior is also common prior knowledge, which assumes that regions with high feature contrast values are more likely to be salient region. e.g., Cheng *et al.* [13] compute different color attributes' global contrast information to represent saliency cues. In [14], the saliency value of each region is determined by the contrast between its feature and image's mean feature. In addition, center prior defines that image center regions are more likely to be salient object because human eye tends to detect center regions instead of surrounding regions.

Recently, machine learning algorithms are widely applied to bottom-up strategies, we call them MLBU methods, which contain two stages: For an input image, prior knowledge is firstly utilized to select some regions as training samples and then label them. Secondly, the selected training samples are used to learn a saliency classifier based on various machine learning algorithms, such as bootstrap learning [4], multi-instance learning [5], dictionary learning [6] and so on. The learned saliency classifier can classify salient object and background so that a saliency map is generated. Nevertheless, a vital drawback about MLBU strategy is that a subset of training set obtained by prior knowledge might be mislabeled in some complex scenes, because prior knowledge loses easily effectiveness when image content is very complex and rich. As a result, unreliable training set is hard to make subsequent learning process succeed.

Inaccurate supervised learning, which is an important branch of weakly supervised learning idea, concerns the situation in which the supervision information is not always ground-truth; In other words, label information might suffer from some errors. Thereby inaccurate supervised learning aims to construct various rules to check the label reliability of each instance. e.g., Muhlenbach *et al.* [36] propose a data-editing approach to check label noises. Zhou [37] aim to infer the ground-truth label based on ensemble strategy. Considering that this situation often occurs in real world, inaccurate supervised learning framework is widely applied to many real tasks, such as image classification [38], deep learning [39] and so forth.

Following the above analysis, we notice that there is a strong relationship between inaccurate supervised learning framework and MLBU methods. Because training labels in MLBU rely on prior knowledge, which may lead to noisy labels in some complex scenes. Thus, the goal of our work is to construct various rules to check noisy labels and improve the quality of training set, which can make subsequent

learning process succeed. To the best of our knowledge, this setting is of great importance but not studied in previous MLBU methods, even all bottom-up methods. Furthermore, we also develop a novel iterative feedback framework and introduce the proposed inaccurate supervised learning process into this framework for better performance achievement.

III. THE PROPSOED METHOD

In this section, the proposed method will be detailed. Firstly, in SecIII-A, we detail feature descriptor which is used to describe each image region in our framework. Then, the proposed inaccurate supervised learning framework is introduced in SecIII-B. Furthermore, we describe the content of iterative feedback framework in SecIII-C. Finally, smoothness operator is utilized in SecIII-D.

A. FEATURE DESCRIPTOR

Given an input image I, the SLIC algorithm [15] is firstly utilized to segment it into N superpixels as basic units, which are defined as $S = \{s_1, s_2, \ldots, s_N$. Inspired by [9], we use pre-trained VGG19 net [16] to extract multiple deep feature maps from input image to describe its content, which can better exploit the high-level semantic information of image. As suggested in [9], the first and last layer of VGG19 net is utilized to extract deep feature maps from input image. The total number of the extracted deep feature maps about input image I is set to M, so that superpixel s_i $(i = 1, 2, \ldots, N)$ is represented by a M-dimensional deep feature vector, where m-th component is the value of superpixel s_i in m-th deep feature map. We define the deep feature of superpixel s_i to be d_i in our method.

B. INACCURATE SUPERVISED LEARNING

For input image I, we aim to learn a saliency classifier to classify salient object and background by applying machine learning algorithms. Generally, training set construction about input image plays an important role in learning process. Similar to previous MLBU methods, prior knowledge is firstly used to provide a coarse indicator for training set construction. In our framework, three prior maps containing background-based map, objectness map and global contrast map are constructed, we compute them as follows:

Background-based map: Background prior defines image boundary superpixels to be background seeds, then the saliency value of each superpixel is determined by its feature contrast with the background seeds. Thus, the background-based map $B = [B_1, B_2, \ldots, B_N]^T$ is constructed as follows:

$$B_i = \frac{1}{nb} \sum_{j=1}^{nb} \exp\left(\frac{\left\| d_i, d_j^b \right\|}{\theta}\right) \tag{1}$$

where B_i is the background-based value of superpixel s_i, s_j^b is j-th boundary superpixel. nb is the total number of boundary superpixels. d_i and d_j^b are the deep features of superpixel s_i and boundary superpixel s_j^b, θ is set to 0.1.

Global contrast map: Global contrast prior assumes that superpixels with high feature contrast are more likely to be salient object. Thus, the global contrast map $G = [G_1, G_2, \ldots, G_N]^T$ is constructed as follows:

$$G_i = \frac{1}{N} \sum_{j=1}^{N} \exp\left(\frac{\left\| d_i, d_j \right\|}{\theta}\right) \tag{2}$$

where G_i is the global contrast value of superpixel s_i. N is the number of superpixels. d_i and d_j are the deep features of superpixel s_i and s_j, θ is set to 0.1.

Objectness map: Objectness prior is proposed in [31], it uses five prior cues to compute the likelihood of the given window containing salient object. Five prior cues contain multi-scale saliency, color-contrast, edge density, superpixel straddling and location plus size. Here, we generate directly Objectness map without any modifications, which is defined as $O = [O_1, O_2, \ldots, O_N]^T$, where O_i is the objectness value of superpixel s_i.

The visual results of three prior maps are shown in Fig.2(b)-(d). Then the coarse saliency map U is constructed by integrating three prior maps, i.e., $U = B + O + G$; Fig.2(e) shows the visual result of the coarse saliency map integrating three prior maps, we can see that it can further improve performance than any prior map. Based on the coarse saliency map U, we can select some superpixels as training samples by setting appropriate threshold. i.e., superpixel s_i is selected as positive sample (foreground) if its coarse saliency value is higher than $th + \beta$, and superpixel s_i is selected as negative sample (background) if its coarse saliency value is lower than $th - \beta$, where th is the mean value of the coarse saliency map and parameter β is set to 0.3 in our method. As a result, a coarse training set is constructed, Fig.2(f) shows the result of the coarse training set, superpixels covered by red and blue regions are respectively positive samples and negative samples. For the first image, it's observed that an accurate training set is generated. While we also find that there are some noisy labels in the coarse training set for some images, such as the second and third image, these inaccurate training samples fail to make subsequent learning process succeed.

Considering that a subset of training set might be mislabeled due to the limitation of prior knowledge. We propose to construct various rules to find these inaccurate samples which might be mislabeled by prior knowledge. To the best of our knowledge, this setting is rarely studied but of great importance in MLBU methods. To achieve this goal, we propose three checking rules to check the label reliability of each training sample and remove mislabeled samples from the coarse training set. It's noticed that the label values of positive sample and negative sample are respectively set to 1 and 0 in subsequent computation.

1) LOCAL CONSISTENCY RULE

The local relationship between adjacent superpixels is still an important factor in saliency cues exploitation. Inspired by

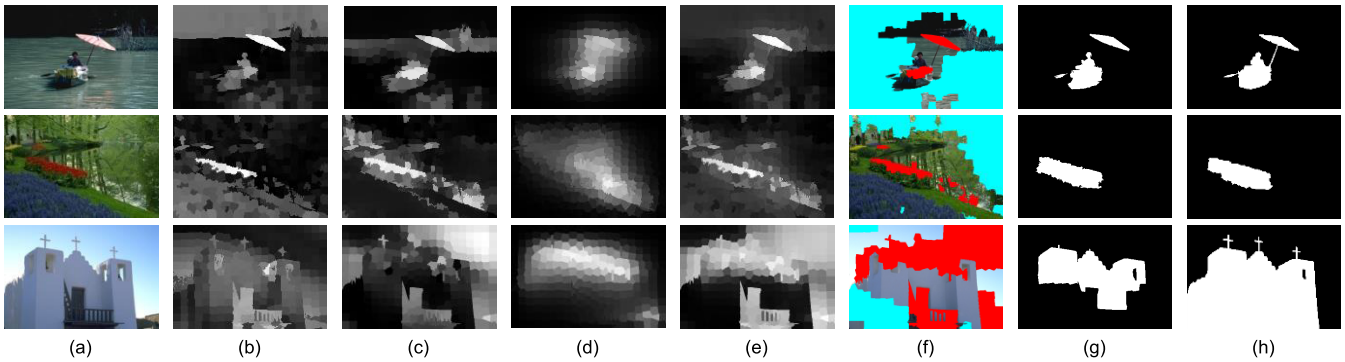

FIGURE 2. The visual results of various steps (a)Image (b)Global contrast map (c)Background-based map (d)Objectness map (e) Coarse saliency map (f)Coarse training set, red and blue regions refer to positive samples and negative samples (g)Saliency result obtained by ISL (h)Ground Truth.

FIGURE 3. The visual result of three checking rules in the ISL (a)Input image (b)Coarse training set obtained by prior knowledge (c) Refined training set obtained by three checking rules.

graph theory [11], two adjacent superpixels are more likely to be assigned to similar saliency values. In other words, there is a label consistency between adjacent samples. Thus, a training sample's label might be inaccurate if its label is different from its most adjacent samples' labels, we name this kind of sample as "isolated sample", whose label is usually hard to be detected correctly, such as sample covered by yellow box in Fig.3(b). Thus, a sample will be removed from the coarse training set if its most adjacent samples' labels are opposite with it. More specifically, given a sample r_i, it is introduced into the following equation:

$$\frac{1}{n_{r_i}} \sum_{r_j \in adj(r_i)} \left| L\left(r_i\right) - L\left(r_j\right) \right| > th1 \tag{3}$$

where $L\left(r_i\right)$ is the label value of r_i and $adj(r_i)$ is the set of adjacent training samples of r_i, r_j is the adjacent training sample of r_i and $L\left(r_j\right)$ is its label value. n_{r_i} is the total number of the adjacent samples of r_i. $th1$ is set to 0.8. $\left|L\left(r_i\right) - L\left(r_j\right)\right| = 0$ if they have the same label, 1 otherwise. Equation (3) means that more than 80% of the adjacent samples have opposite labels with sample r_i, thus, sample r_i will be defined as "isolated sample", we will remove it from training set, because it is more likely to be inaccurate sample.

2) FEATURE CONTRAST RULE

Generally, there is a feature difference between salient object and background, this kind of feature difference is still existed even if salient object and background have similar features. Thus, we can infer that there is also feature difference between positive samples set and negative samples set. In other words, a sample should share more similar features with samples which have the same label with it. Based on above observation, given a training sample r_i, it is introduced into the following equation

$$\frac{\frac{1}{na} \sum\limits_{r_j \in SC} \exp(-\frac{\|c_i - c_j\|}{\theta})}{\frac{1}{nb} \sum\limits_{r_j \in DC} \exp(-\frac{\|c_i - c_j\|}{\theta})} < 1 \tag{4}$$

where SC is the set of samples which have the same label with sample r_i. DC is the set of samples which have the opposite label with sample r_i. c_i and c_j are the deep features of sample r_i and r_j. na and nb are the total number of samples in the set SC and DC, parameter θ is set to 0.1. Training sample r_i is more similar with samples in the set DC if it satisfies equation(4), i.e., it is more similar with samples which have opposite label with it. Thus, we consider that sample r_i might be inaccurate and remove it from the coarse training set.

3) SPATIAL COMPACTNESS RULE

Different from the first two rules, spatial compactness rule focus on the label reliabilities of all positive samples. Generally, salient object has compactness spatial structure instead of wide spatial distribution. It means that positive samples tend to be clustered for each other in most images. Thus, a positive sample is considered to be unreliable if it is far from most positive samples, such as samples covered by black boxes in Fig3(b), they will be removed from the coarse training set. Specifically, given a positive sample r_i, it is introduced into the following equation:

$$\|p\left(r_i\right) - w\_center\| > th2 \tag{5}$$

where

$$w\_center = \frac{1}{Np} \sum\nolimits_{j=1}^{Np} \left(U(r_j) \times p(r_j)\right) \tag{6}$$

In equation(5), r_i is i-th positive sample and $p\left(r_i\right)$ is the position coordinate of r_i (The position coordinate of a sample is the mean position coordinate of pixels within this sample),

w_center is the weighted position center of all positive samples, which is computed in equation(6), where r_j is j-th positive sample, $p(r_j)$ is the position coordinate of sample r_j. $U(r_j)$ is the coarse saliency value of r_j and Np is the total number of positive samples. It's noticed that the weighted position center *w_center* is more likely to be the center of salient object, such as the green point in Fig.3(b). Instead of computing directly the mean position center of all positive samples, each positive sample has a weight in the weighted position center computation. Because positive samples are defined as superpixels whose coarse saliency values are higher than a unified threshold $th + \beta$ in our method, i.e., various positive samples might have different coarse saliency values even if they all have positive labels. We consider that the weighted position center tends to near positive samples with high coarse saliency values, because they are more reliable. Based on observation that salient object usually has compactness spatial structure, positive sample r_i will be removed from the coarse training set if it satisfies equation(5), because it is far from the weighted position center which is more likely to near most reliable positive samples, such as samples covered by black boxes in Fig.3(b). They are more likely to be inaccurate samples.

In summary, given a training sample r_i, above three rules are utilized to check its label reliability. For one thing, sample r_i will be introduced into all three rules if it is positive sample. We consider that positive sample r_i might be inaccurate as long as it satisfies any one of three rules. Thus, there is not the priority between three rules. For the other, sample r_i will be introduced into the first two rules if it is negative sample. Also, negative sample r_i might be inaccurate as long as it satisfies any one of the first two rules. After removing all inaccurate samples from the coarse training set, so that an optimized training set is generated.

Based on the optimized training set, we aim to learn a saliency classifier to predict each superpixel to be foreground/background by applying machine learning algorithms. Here, simple support vector machine (SVM) is utilized due to its effectiveness in classification problem. Based on the optimized training set, SVM based saliency classifier is learned as follows:

$$\begin{aligned} &\min_{w,z} \frac{1}{2}\|w\|^2 \\ &\text{s.t. } L(r_i) \times \left(w^T c_i + z\right) \geq 1, \quad i = 1, \ldots, n \end{aligned} \tag{7}$$

where $L(r_i)$ is the label value of sample r_i, c_i is the feature of sample r_i. w and z are the parameters of saliency classifier. Actually, equation(7) is the classical SVM solution formula, the parameter w and z of saliency classifier can be obtained by solving above function, then saliency classifier is generated to predict each superpixel. As a result, we can obtain a binary saliency map, in which each superpixel's saliency value is 1(foreground) or 0(background). The visual result of the proposed ISL is shown in Fig.2(g).

C. ITERATIVE FEEDBACK CLASSIFICATION FRAMEWORK

The proposed ISL can achieve better saliency result by refining the quality of the coarse training set. However, the ISL still rely on the performance of the coarse training set obtained by prior knowledge. In other words, the ISL can further refine saliency result on the premise that the coarse training set is coarse and imprecise (i.e., most samples are accurate, while a small subset is mislabeled, such as the second image in Fig.2). However, the result of ISL is far from satisfaction when most samples in the coarse training set all indicate false labels (such as the third image in Fig.2). Based on above observation, different from conventional classification framework using directly training set to learn a saliency classifier to implement classification task. We study a novel iterative feedback (IF) framework and introduce the ISL process into the IF framework for more accurate saliency result.

We list the framework of the IFISL in **Algorithm.1.** The IFISL is an iterative mechanism, the ISL process is implemented at each iteration. i.e., an iteration contains four stages: the coarse training set construction, mislabeled samples remove, saliency classifier construction, saliency result generation. The coarse training set at $t+1$-th iteration is self adjusted according to the feedback of saliency result after t-th iteration. So that an iterative feedback mechanism is constructed, in which saliency result is optimized gradually to an optimal state.

Next, the construction of the coarse training set at $t+1$-th iteration in the IFISL is detailed, we firstly define the coarse saliency map at $t+1$-th iteration to be U^{t+1}, which is computed as follows:

$$U^{t+1} = \sum_{P \subset \{B,C,G\}} RS(P) \times P \tag{8}$$

where

$$RS(P) = 1 - \frac{1}{N}\sum_{j=1}^{N} \exp\left(\left|P_i - S_i^t\right|\right) \tag{9}$$

where the coarse saliency map at $t+1$-th iteration U^{t+1} is the weight summation of three prior maps. P refers to prior map, i.e., $P \subset \{B, O, G\}$, where B is the background-based map, O is the objectness map and G is the global contrast map. $RS(P)$ is the weight of prior map P, it is computed by equation(9), where S^t is the saliency result after t-th iteration and S_i^t is the value of superpixel s_i in S^t. P_i is the value of superpixel s_i in prior map P, N is the total number of superpixels. Actually, higher $RS(P)$ indicates that prior map P and saliency result after t-th iteration S^t are more similar, thus, we will improve its weight in the construction of the coarse saliency map at $t+1$-th iteration (i.e., U^{t+1}). It's noticed that the weight of each prior map is 1/3 in the initial coarse saliency map construction, i.e., when $t = 1$. By setting adaptive threshold to the coarse saliency map at $t+1$-th iteration (it is defined

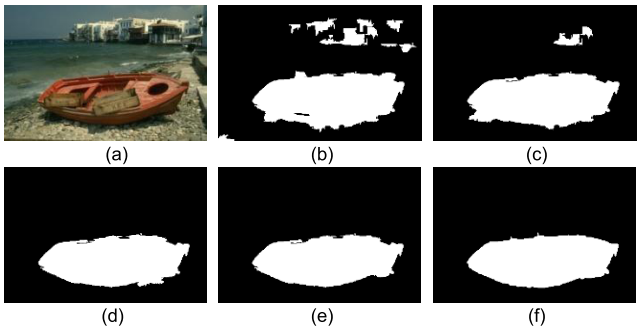

FIGURE 4. Saliency results after different iteration times in the IFISL (a)Image (b) T = 1 (c) T = 2 (d) T = 3 (e) T = 4 (f) Ground Truth.

in the first paragraph of SecIII-B), we can obtain the coarse training set at $t+1$-th iteration.

As a result, the coarse training set will be gradually refined by self adjusting the weight of three prior knowledge according to the feedback. While three checking rules are also utilized after the coarse training set construction of each iteration for more accurate training set acquirement. So that saliency classifier can be optimized gradually to generate stable saliency result. Based on experimental analysis, iteration times T is set to 4 in our method. Fig.4 shows saliency results after different iteration times. We can see that saliency result is optimized gradually to optimal state with iteration. It's noticed that saliency result after the first iteration corresponds to saliency result obtained by ISL without IF. We can find that it is further refined by the proposed IF framework.

Algorithm 1 The Flow of the Proposed IFISL

Initialization: $t = 1, T = 4$
For $t \leq T$
1: **if** $t = 1$
2: Initialize U^t by giving the same weight to each prior map
3: **else**
4: Compute U^t according to equation(8)
5: **end**
6: Obtain the coarse training set R^t based on U^t
7:Obtain the optimized training set R_o^t according to the proposed three rules
8: Based on R_o^t, learn saliency classifier according to equation(7)
9: Obtain saliency map S^t, in which each superpixel is predicted to be foreground/background by learned saliency classifier.
10: $t = t+1$
end
Output: S^T

D. SMOOTHNESS OPERATOR

Finally, we use superpixel-level smoothness function to further smooth saliency map obtained by IFISL. The affinity matrix $W \epsilon \mathbb{R}^{N \times N}$ is firstly constructed to exploit the relationship between superpixels, where element w_{ij} is computed as follows:

$$w_{ij} = \begin{cases} \exp\left(-\frac{\|d_i - d_j\|}{\theta}\right) & \text{if } s_j \epsilon adj(s_j) \\ 0 & \text{otherwise} \end{cases} \quad (10)$$

where w_{ij} is the weight between superpixel s_i and s_j if they are neighbors for each other, 0 otherwise. $adj(s_i)$ is the set of neighbors of superpixel s_i, d_i and d_j are the deep features of superpixel s_i and s_j, parameter θ is set to 0.1.

Then, saliency map obtained by IFISL is defined to be $X = [x_1, x_2, \ldots, x_N]^T$, where x_i is the value of superpixel s_i in saliency map X. Then, smoothness saliency map $Y = [y_1, y_2, \ldots, y_N]^T$ is computed according to $Y = WX$, where y_i is the value of superpixel s_i in the final smoothness saliency map Y. i.e., the saliency value of each superpixel is determined by the weight summation of its adjacent superpixels' saliency values in smoothness operator. The visual result of smoothness operator is shown in Fig.5, it's observed that salient regions are better highlighted and background noises are further suppressed by smoothness operator.

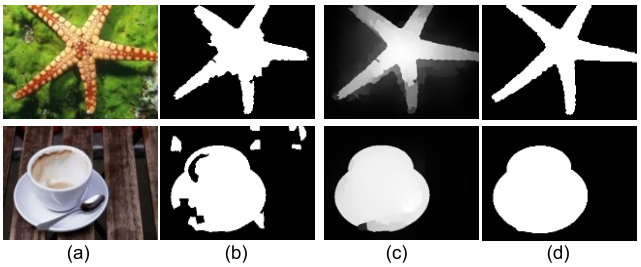

FIGURE 5. The visual result of smoothness operator (a) Input image (b) Saliency map obtained by IFISL (c)Saliency map obtained by smoothness operator (d) Ground Truth.

IV. EXPERIMENTS

The proposed method is compared with other 13 state-of-the-art methods, including MLSP[17], TLLT[18], BSCA[19], LDS[20], MST[21], MAP[22], AE[23], SMD[24], HCA[9], LEGs[7], S-CNN[8], FCB[25] and DGLS[6]. Where LDS, MST and DGLS are machine learning based bottom-up (MLBU) methods. HCA, LEGs, S-CNN and AE exploit saliency cues by utilizing CNN based deep learning framework. MLSP, TLLT, BSCA and MAP belong to graph-based optimization methods. In addition, FCB and SMD also achieve outstanding performances in recent years. All saliency maps are obtained by running codes or directly provided by authors. It's noticed that we only can obtain saliency maps about HCA, AE, LEGs and S-CNN on several datasets instead of all datasets, because authors only provide their results on several datasets.

All methods are compared on three benchmark datasets, including ECSSD[26], DUT-OMRON[11] and SOD[27]. More specifically, ECSSD dataset is composed of 1000 images, most of which are natural scenes, such as people, animal, tree and so on. SOD dataset is composed of

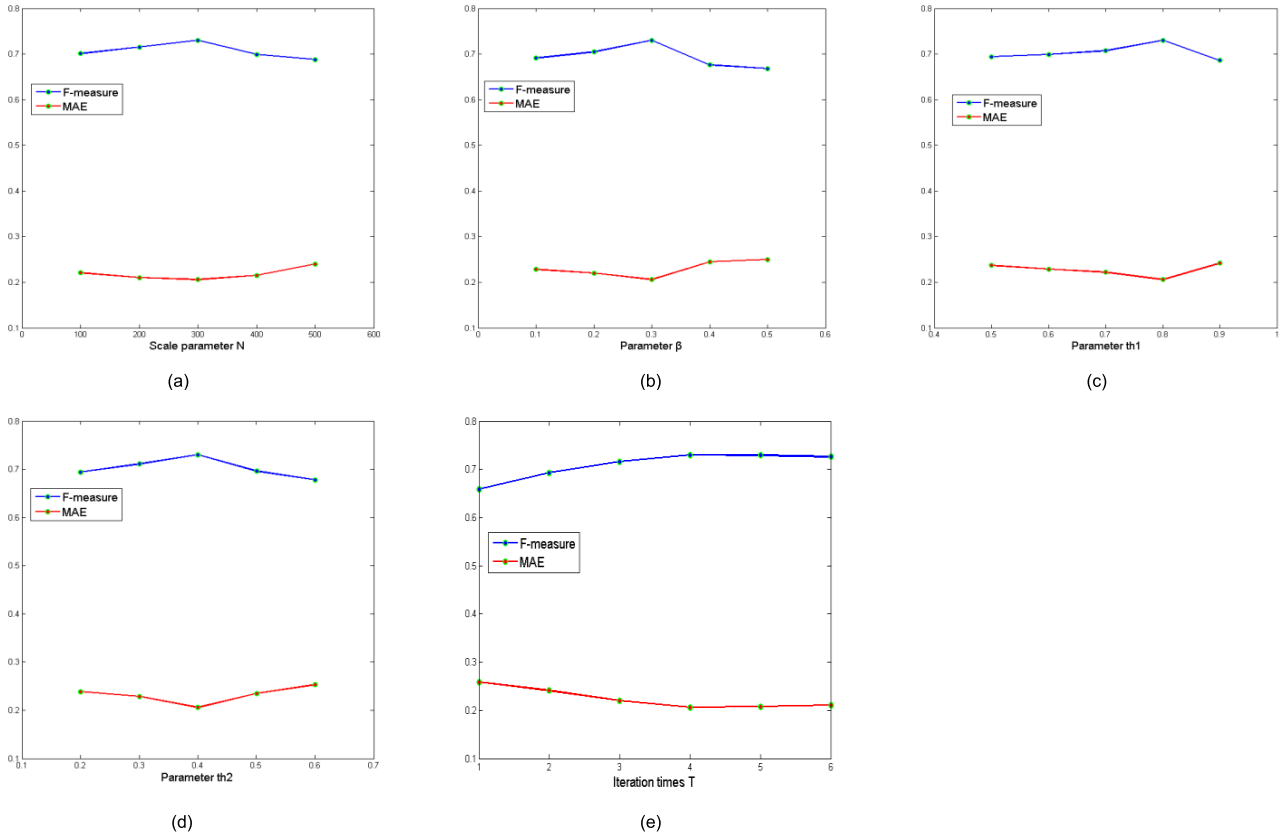

FIGURE 6. **Parameters analysis (a) Scale parameter N analysis (b) Parameter β analysis (c) Parameter th1 analysis (d) Parameter th2 analysis (e)Iterative times T analysis.**

300 images and more complex than ECSSD, salient object and background are hard to be separated in most images of SOD dataset. In contrast, DUT-OMRON dataset contains 5172 images, which have rich semantic information, existing works are hard to achieve top performances on the DUT-OMRON dataset. Above three datasets are very classical in saliency detection field.

There are three evaluation metrics in our experiments, including Precision-Recall (PR) curve, F-measure score and Mean Absolute Error (MAE) value. PR curve is obtained by comparing the ground truth and binary map using different thresholds from 0 to 255 to segment saliency map. F-measure score is an overall evaluation metric incorporating precision rate and recall rate. As supplement, MAE value is also introduced into our experiments, it evaluates performance by computing directly the mean difference between saliency map and the ground truth.

A. PARAMETERS ANALYSIS

In order to detect intuitively the performances of the proposed method when various parameters are set to different values. Two quantitative metrics are used to evaluate performance, i.e., F-measure score and MAE value. In addition, we only test parameters sensitiveness experiments on the SOD dataset, then the best parameters are applied to other datasets. Experimental results are shown in Fig.6. Firstly, our method is superpixel-level method, thus, scale parameter N is very important for the final result. The performance of the proposed method when scale parameter N is set to different values is shown in Fig.6(a), it's observed that the best N is set to 300. Secondly, parameter β is key parameter in the coarse training set construction, we can find that the best performance is obtained when parameter β is set to 0.3 by observing Fig.6(b). In addition, Fig.6(c) shows the analysis result of threshold parameter $th1$, it's observed that the performance when $th1 = 0.8$ is superior obviously to other values. Furthermore, it's surprised to find that the best threshold parameter $th2$ is set to 0.4 in Fig.6(d). Finally, we test the best iteration times of the proposed method, which play an important role for the final performance improvement. We can see that the iteration times T is set to 4 in Fig.6(e). It's noticed that the proposed method when $T = 1$ is an inaccurate supervised learning framework without iteration feedback mechanism, we are surprised to find that it makes a great contribution for the final performance improvement.

B. ABLATION STUDY

To validate the effectiveness of each step in our method, ablation experiments are tested, it's noticed that the selection of evaluation metric and dataset is the same with parameter analysis experiments. Quantitative results are shown in Fig.7.

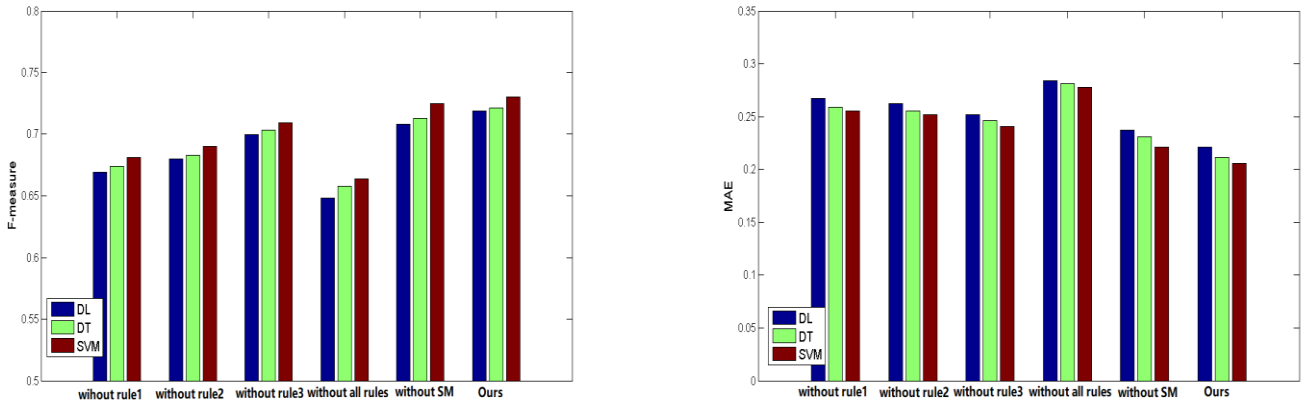

FIGURE 7. The result of ablation experiment, "rule1" is local consistency rule, "rule2" is feature contrast rule and "rule3" is spatial compactness rule. "SM" refers to smoothness operator. 'DL', 'DT' and 'SVM' represent respectively Dictionary Learning, Decision Tree and Support Vector Machine.

Firstly, the proposed method considers that training labels obtained by prior knowledge might be inaccurate and focuses on constructing three checking rules to check the label reliabilities of training samples then remove these samples with noisy labels, so that a more accurate training set is generated to make subsequent learning process succeed. Two figures in Fig.7 show that three checking rules all make contributions for the final performance improvement, especially the local consistency rule and feature contrast rule (rule1 is local consistency rule, rule2 is feature contrast rule and rule3 is spatial compactness rule). Also, ''without all three rules'' indicates that there is not label checking process at each iteration while others are unchanged. We can find that three proposed checking rules play important roles in the whole framework, it demonstrates the rationality of the proposed ISL. Furthermore, we are surprised to find that the final smoothness operator also contributes to the final performance improvement. In our method, the classical SVM algorithm is used in the construction of saliency classifier. Furthermore, the influence of saliency classifier selection is also validated, various saliency classifiers are utilized in our framework, including dictionary learning (DL), decision tree (DT) and support vector machine (SVM). i.e., we use DL or DT to replace SVM to construct saliency classifier while other components are unchanged in our framework. Experimental result is also shown in Fig.7, the performance using SVM is superior obviously to DL and DT on the premise that the other components of the proposed method are invariable. As seen from Fig.7, we also find that the influences of three proposed checking rules are larger obviously than subsequent saliency classifier selection (i.e., machine learning algorithms selection) in the whole framework. This demonstrates the necessity of the proposed method.

C. TRAINING SET SIZE AND OVERFITTING PROBLEM

We do not set the size of training set in the proposed framework. Instead, training set is constructed by setting adaptive threshold, because the number of reliable samples we can obtain might be different in various images; e.g., more training samples can be generated for the first image in Fig.2, in contrast, we only can obtain relatively less training samples in the second image of Fig.2. Generally, the size of training set might be determined by salient object scale, image content complexity and so on. Thus, for MLBU methods, it's hard to set a fixed training set size for an image in advance. In addition, overfitting problem also needs to be analyzed due to the existence of learning process. Although we do not set a fixed training set size for each image, based on experimental analysis, we find that the proportion of training samples per image is about 55%, which is a reasonable range for avoiding the overfitting problem of learning process. Furthermore, three proposed checking rules can reduce effectively the noisy labels of training set by exploiting the relationship between samples. It's no doubt that this operator also can further avoid the overfitting problem.

D. COMPARISON EXPERIMENTS

The comparison results of all methods about F-measure score and MAE value are listed in Table.1 (It's noticed that lower MAE value indicates better performance and higher F-measure score refers to better performance. In addition, bold words represent the best values). It's seen from Table.1, it's surprised to find that the proposed method achieves the best performance on all datasets, especially on the DUT-OMRON dataset, superiority is the largest. As MLBU methods, LDS, MST and DGLS are all inferior to the proposed method, it illustrates the effectiveness of the proposed iterative feedback based inaccurate supervised learning framework (IFISL). Furthermore, we also find that the proposed method also outperforms AE, HCA, LEGs and S-CNN, which are outstanding deep neural network (DNN) based methods. Furthermore, Fig.8 shows the comparison results of all methods about PR curve, we can see that the proposed method is superior to other methods, especially

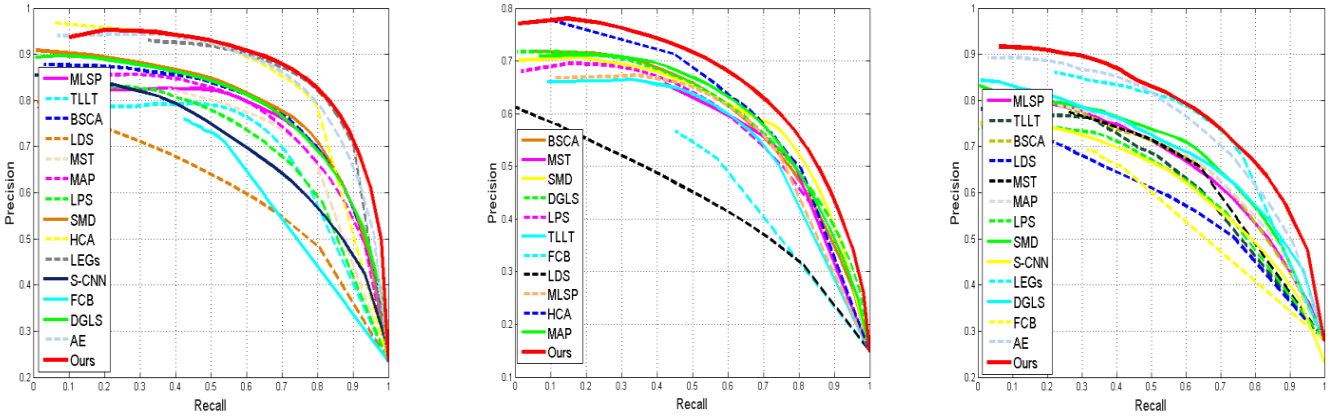

FIGURE 8. The PR curve comparison results of all methods on three datasets (a) ECSSD (b)DUT-OMRON (c)SOD.

TABLE 1. The quantitative comparison results on three datasets (Bold words refer to the best values, '-' indicates that we cannot obtain the saliency results of corresponding methods).

	ECSSD		DUT-OMRON		SOD	
	F-measure	MAE	F-measure	MAE	F-measure	MAE
MLSP	0.731	0.205	0.551	0.183	0.631	0.262
TLLT	0.718	0.177	0.589	0.141	0.622	0.240
BSCA	0.736	0.182	0.530	0.191	0.638	0.251
LDS	0.596	0.221	0.432	0.174	0.572	0.262
MST	0.726	0.148	0.551	0.149	0.649	0.222
MAP	0.734	0.185	0.543	0.183	0.641	0.252
AE	0.821	0.169	-	-	0.713	0.234
SMD	0.747	0.171	0.529	0.173	0.663	0.234
HCA	0.808	0.119	0.561	0.156	-	-
LEGs	0.809	0.118	0.544	0.145	0.724	0.211
S-CNN	0.690	0.214	-	-	0.596	0.217
FCB	0.701	0.130	0.528	0.149	0.546	0.244
DGLS	0.743	0.187	0.529	0.178	0.644	0.247
Ours	**0.828**	**0.115**	**0.591**	**0.135**	**0.730**	**0.206**

several MLBU methods. Also, AE and HCA have outstanding performances, which are still inferior to our method. Three datasets are composed of numerous complex scenes, some of which have rich semantic information, top performances demonstrate the superiority of the proposed method.

Fig.9 shows the visual results of all methods on some example images, it's surprised to find that the proposed method can capture effectively the contrast between salient object and background. For the first two images which have rich features, it's observed that the proposed method can highlight completely salient region, the each component of salient object can be detected even if it is composed of multiple regions with different features. For the third image where salient object and background share similar features, the proposed method still can separate correctly salient object and background. Considering that the background prior is utilized in the coarse saliency map construction, we also show the performance of the proposed method when salient object touches image boundary in the fourth image, a good result is still generated by our method. We can infer that foreground superpixels touching image boundary are still highlighted by our method even if they are mistaken for background in the coarse saliency map. Furthermore, the last image is a very difficult image for existing works, where a part of salient object is very similar with background while salient object has rich features. It's surprised to find that the proposed method still can detect correctly salient object. For above images, other comparison methods lose easily effectiveness, some even mark falsely salient object. Visual comparison results further validate the superiority of the proposed method.

E. RUNTIME ANALYSIS

Runtime analysis experiments are tested on the SOD dataset for convenience. The average running time per image on

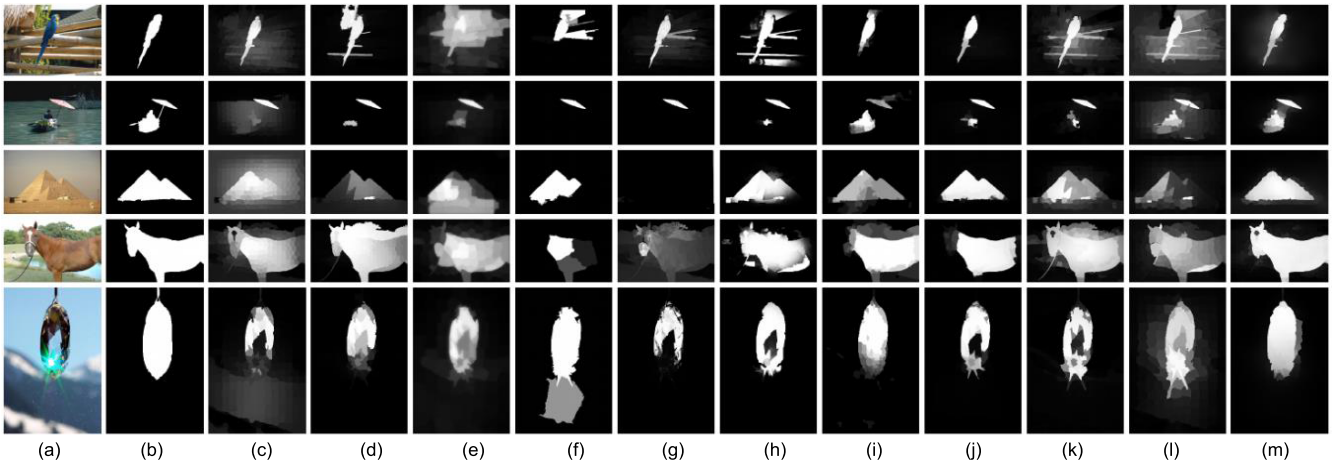

FIGURE 9. The visual comparison results on several images. (a)Input image (b)Ground Truth (c)BSCA (d) TLLT (e) DGLS (f) FCB (g) LPS (h) MST (i) LEGs (j)HCA (k)SMD (l)MLSP (m)Ours.

the SOD dataset is 1.54s. Actually, deep feature extraction spends averagely 0.79s on each image, which is the most time-consuming work. Four iterations are implemented in our method, each iteration in the IFISL spends averagely 0.17s. Finally, 0.07s is spent by smoothness operator. In summary, the proposed method achieves outstanding performance in efficiency.

V. CONCLUSIONS

Based on observation that training set construction in existing MLBU methods usually rely on prior knowledge, which loses easily effectiveness in some complex scenes. In this paper, we present an iterative feedback based inaccurate supervised learning framework for saliency detection. Given an input image, an inaccurate supervised learning (ISL) framework is proposed, which contains the coarse training set construction, label reliability checking, saliency classifier construction and saliency result prediction. Comparing with previous works, the proposed checking rules can effectively refine the coarse training set and make subsequent learning process succeed, so that a better saliency map is generated by our ISL. Furthermore, we introduce the ISL into a novel iterative feedback (IF) framework. At each iteration, an ISL process is implemented, and the coarse training set at certain iteration is self adjusted according to the feedback of saliency result after last iteration. As a result, saliency result is optimized gradually to a stable state with iteration. Finally, smoothness operator is also utilized to further smooth saliency map. Experimental results on three datasets demonstrate adequately the superiority of the proposed method.

REFERENCES

[1] J. Huang, X. Yang, X. Fang, W. Lin, and R. Zhang, "Integrating visual saliency and consistency for re-ranking image search results," *IEEE Trans. Multimedia*, vol. 13, no. 4, pp. 653–661, Aug. 2011.

[2] C. Luo, L. Jin, and Z. Sun, "MORAN: A multi-object rectified attention network for scene text recognition," *Pattern Recognit.*, vol. 90, pp. 109–118, Jun. 2019.

[3] X. Zhao, H. Wang, J. Wu, Y. Li, and S. Zhao, "Remote sensing image segmentation using geodesic-kernel functions and multi-feature spaces," *Pattern Recognit.*, vol. 104, Aug. 2020, Art. no. 107333.

[4] H. Lu, X. Zhang, J. Qi, N. Tong, X. Ruan, and M.-H. Yang, "Co-bootstrapping saliency," *IEEE Trans. Image Process.*, vol. 26, no. 1, pp. 414–425, Jan. 2017.

[5] F. Huang, J. Qi, H. Lu, L. Zhang, and X. Ruan, "Salient object detection via multiple instance learning," *IEEE Trans. Image Process.*, vol. 26, no. 4, pp. 1911–1922, Apr. 2017.

[6] M. Zhang, Y. Wu, Y. Du, L. Fang, and Y. Pang, "Saliency detection integrating global and local information," *J. Vis. Commun. Image Represent.*, vol. 53, pp. 215–223, May 2018.

[7] L. Wang, H. Lu, X. Ruan, and M.-H. Yang, "Deep networks for saliency detection via local estimation and global search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3183–3192.

[8] S. He, R. Lau, and W. Liu, "SuperCNN: A superpixel wise convolutional neural network for salient object detection," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 330–334, Apr. 2015.

[9] Y. Qin, M. Feng, H. Lu, and G. W. Cottrell, "Hierarchical cellular automata for visual saliency," *Int. J. Comput. Vis.*, vol. 126, no. 7, pp. 752–770, Jul. 2018.

[10] Y. Zeng, M. Feng, H. Lu, G. Yang, and A. Borji, "An unsupervised game-theoretic approach to saliency detection," *IEEE Trans. Image Process.*, vol. 27, no. 9, pp. 4545–4554, Sep. 2018.

[11] L. Zhang, C. Yang, H. Lu, R. Xiang, and M.-H. Yang, "Ranking saliency," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 9, pp. 1892–1904, Sep. 2017.

[12] H. Li, H. Lu, Z. Lin, X. Shen, and B. Price, "Inner and inter label propagation: Salient object detection in the wild," *IEEE Trans. Image Process.*, vol. 24, no. 10, pp. 3167–3186, Oct. 2015.

[13] M. Cheng, N. J. Miltra, X. Huang, P. H. S. Torr, and S. Hu, "Global contrast based salient region detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 569–582, Mar. 2015.

[14] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1597–1604.

[15] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, "SLIC superpixels compared to state-of-the-art superpixels methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012.

[16] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, May 2015, pp. 1–14.

[17] I. Hwang, S. H. Lee, J. S. Park, and N. I. Cho, "Saliency detection based on seed propagation in a multilayer graph," *Multimedia Tools Appl.*, vol. 76, no. 2, pp. 2111–2129, Jan. 2017.

[18] C. Gong, D. Tao, W. Liu, S. J. Maybank, M. Fang, K. Fu, and J. Yang, "Saliency propagation from simple to difficult," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2531–2539.

[19] Y. Qin, H. Lu, Y. Xu, and H. Wang, "Saliency detection via cellular automata," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 110–119.
[20] S. Fang, J. Li, Y. Tian, T. Huang, and X. Chen, "Learning discriminative subspaces on random contrasts for image saliency analysis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 5, pp. 1095–1108, May 2017.
[21] W.-C. Tu, S. He, Q. Yang, and S.-Y. Chien, "Real-time salient object detection with a minimum spanning tree," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2334–2342.
[22] J. Sun, H. Lu, and X. Liu, "Saliency region detection based on Markov absorption probabilities," *IEEE Trans. Image Process.*, vol. 24, no. 5, pp. 1639–1649, May 2015.
[23] L. Zhang, J. Ai, B. Jiang, H. Lu, and X. Li, "Saliency detection via absorbing Markov chain with learnt transition probability," *IEEE Trans. Image Process.*, vol. 27, no. 2, pp. 987–998, Feb. 2018.
[24] H. Peng, B. Li, H. Ling, W. Hu, W. Xiong, and S. J. Maybank, "Salient object detection via structured matrix decomposition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 818–832, Apr. 2017.
[25] G.-H. Liu and J.-Y. Yang, "Exploiting color volume and color difference for salient region detection," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 6–16, Jan. 2019.
[26] Q. Yan, L. Xu, J. Shi, and J. Jia, "Hierarchical saliency detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1155–1162.
[27] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille, "The secrets of salient object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 270–290.
[28] Y. Pang, X. Yu, Y. Wang, and C. Wu, "Salient object detection based on novel graph model," *J. Vis. Commun. Image Represent.*, vol. 65, Dec. 2019, Art. no. 102676.
[29] Y. Pang, X. Yu, Y. Wu, and C. Wu, "FSP: A feedback-based saliency propagation method for saliency detection," *J. Electron. Imag.*, vol. 29, no. 1, Jan. 2020, Art. no. 013011.
[30] S. Dai and D. Li, "Research on an infrared multi-target saliency detection algorithm under sky background conditions," *Sensors*, vol. 20, no. 2, Jan. 2020, Art. no. 20020459.
[31] B. Alexe, T. Deselaers, and V. Ferrari, "Measuring the objectness of image windows," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2189–2202, Nov. 2012.
[32] J. Wu, H. Yu, J. Sun, W. Qu, and Z. Cui, "Efficient visual saliency detection via multi-cues," *IEEE Access*, vol. 7, pp. 14728–14735, 2019.
[33] L. Wang, H. Lu, Y. Wang, M. Feng, D. Wang, B. Yin, and X. Ruan, "Learning to detect salient objects with image-level supervision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3796–3805.
[34] M. Qian, J. Qi, L. Zhang, M. Feng, and H. Lu, "Language-aware weak supervision for salient object detection," *Pattern Recognit.*, vol. 96, Dec. 2019, Art. no. 106955.
[35] L. Tang, Q. Wu, W. Li, and Y. Liu, "Deep saliency quality assessment network with joint metric," *IEEE Access*, vol. 6, pp. 913–924, Nov. 2018.
[36] F. Muhlenbach, S. Lallich, and D. A. Zighed, "Identifying and handling mislabelled instances," *J. Intell. Inf. Syst.*, vol. 22, no. 1, pp. 89–109, Jan. 2004.
[37] Z. Zhou, *Ensemble Methods: Foundations and Algorithms*. New York, NY, USA: Taylor & Francis, 2012.
[38] T. Zhou, Y. Li, and G. Gui, "Noise learning based discriminative dictionary learning algorithm for image classification," *J. Franklin Inst.*, vol. 357, no. 4, pp. 2492–2513, Mar. 2020.
[39] D. Hao, L. Zhang, J. Sumkin, A. Mohamed, and S. Wu, "Inaccurate labels in weakly supervised deep learning: Automatic identification and correction and their impact on classification performance," *IEEE J. Biomed. Health Informat.*, early access, Feb. 2020, doi: 10.1109/JBHI.2020.2974425.
[40] Y. Zeng, Y. Zhuge, H. Lu, L. Zhang, M. Qian, and Y. Yu, "Multi-source weak supervision for saliency detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6067–6076.
[41] K.-J. Hsu, Y.-Y. Lin, and Y.-Y. Chuang, "Weakly supervised salient object detection by learning a classifier-driven map generator," *IEEE Trans. Image Process.*, vol. 28, no. 11, pp. 5435–5449, Nov. 2019.

YU PANG received the B.E. degree from Bohai University, in 2015, and the M.S. degree from Northeast Normal University, in 2018. He is currently pursuing the Ph.D. degree with the Faculty of Robot Science and Engineering, Northeastern University. His research interests include computer vision, salient object detection, and video visual tracking.

YUNHE WU received the B.E. degree from Bohai University, in 2016, and the M.S. degree from Northeast Normal University, in 2019. She is currently pursuing the Ph.D. degree with the School of Information Science and Engineering, Northeastern University. Her research interests include computer vision, salient object detection, and stereo matching.

CHENGDONG WU received the B.E. degree from Shenyang Jianzhu University, in 1983, the B.S. degree from Tsinghua University, in 1988, and the Ph.D. degree from Northeastern University, in 1994. He is currently a Full Professor with the Faculty of Robot Science and Engineering, Northeastern University. His research interests include robot vision, computer vision, and medical image processing.

XIAOSHENG YU received the Ph.D. degree from Northeastern University, in 2014. He is currently a Full Lecturer with the Faculty of Robot Science and Engineering, Northeastern University. His research interests include wireless sensor and medical image processing.

YUAN GAO received the M.S. degree in control science and engineering from Shenyang University, in 2016. He is currently pursuing the Ph.D. degree with Northeastern University, Shenyang, China. His research interests include medical imaging processing, machine learning, and pattern recognition.