A Benchmark for Breast Ultrasound Image Segmentation (BUSIS)

Min Xian[†1], Yingtao Zhang[†2], H. D. Cheng*[2,3], Fei Xu[3], Kuan Huang[3], Boyu Zhang[3], Jianrui Ding[4], Chunping Ning[5], Ying Wang[6]

Abstract—**Breast ultrasound (BUS) image segmentation is challenging and critical for BUS Computer-Aided Diagnosis (CAD) systems. Many BUS segmentation approaches have been proposed in the last two decades, but the performances of most approaches have been assessed using relatively small private datasets with different quantitative metrics, which result in discrepancy in performance comparison. Therefore, there is a pressing need for building a benchmark to compare existing methods using a public dataset objectively, and to determine the performance of the best breast tumor segmentation algorithm available today and to investigate what segmentation strategies are valuable in clinical practice and theoretical study. In this work, we will publish a B-mode BUS image segmentation benchmark (BUSIS) with 562 images and compare the performance of five state-of-the-art BUS segmentation methods quantitatively.**

Index Terms—**Breast ultrasound (BUS) images, segmentation, computer-aided diagnosis (CAD) system, benchmark.**

I. INTRODUCTION

BREAST cancer occurs in the highest frequency in women among all cancers, and is also one of the leading cause of cancer death worldwide [1]. The key to reduce the mortality is to find the signs and symptoms of breast cancer at its early stage. In current clinic practice, breast ultrasound (BUS) imaging with computer-aided diagnosis (CAD) system has become one of the most important and effective approaches for breast cancer detection due to its noninvasive, painless, non-radioactive and cost-effective nature. In addition, it is the most suitable approach for large-scale breast cancer screening and diagnosis in low-resource countries and regions.

CAD systems based on B-mode breast ultrasound (BUS) have been developed to overcome the inter- and intra-variabilities of the radiologists' diagnoses, and have demonstrated the ability to improve the diagnosis performance of breast cancer [2]. Automatic BUS segmentation, extracting tumor region from normal tissue regions of BUS image, is a crucial component in a BUS CAD system. It can change the traditional subjective tumor assessments into operator independent, reproducible and accurate tumor region measurements.

Automatic BUS image segmentation study attracted great attention in the last two decades due to clinical demands and its challenging nature, and generated many automatic segmentation algorithms. We can classify existing approaches into *semi-automatic* and *fully automatic* according to "with or without" user interactions in the segmentation process. In most semi-automatic methods, user needs to specify a region of interest (ROI) containing the lesion, a seed in the lesion, or an initial boundary. Fully automatic segmentation is usually considered as a top-down framework which models the knowledge of breast ultrasound and oncology as prior constraints, and needs no user intervention at all. However, it is quite challenging to develop automatic tumor segmentation approaches for BUS images, due to the low image quality caused by speckle noise, low contrast, weak boundary, and artifacts. Furthermore, tumor size, shape and echo strength vary considerably across patients, which prevent the application of strong priors to object features that are important for conventional segmentation methods.

In previous works, all approaches were evaluated by using private datasets and different quantitative metrics (see Table 1) for performance measurements, which make the objective and effective comparisons among the methods impossible. As a consequence, it remains challenging to determine the best performance of the breast tumor segmentation algorithms available today, what segmentation strategies are valuable in clinic practice and study, and what image features are helpful and useful in improving segmentation accuracy and robustness. We will build a BUS image segmentation benchmark which includes a large public BUS image dataset with ground truth, and investigate some objective and quantitative metrics for segmentation performance evaluation.

In this paper, we present a BUS image segmentation benchmark including **562** B-mode BUS images, and compare **five** state-of-the-art BUS segmentation methods by using **seven**

---

[1] the Department of Computer Science, University of Idaho, Idaho Falls, ID 84302 USA.

[2] the School of Computer Science, Harbin Institute of Technology, Harbin, Heilongjiang 150001 China.

[3] the Department of Computer Science, Utah State University, Logan, UT 84322 USA.

[4] the School of Computer Science, Harbin Institute of Technology, Weihai, Shandong 264209 China.

[5] the Department of Ultrasound, The affiliated Hospital of Qingdao University, Qingdao, Shandong 266003 China

[6] the Department of General Surgery, The Second Hospital of Hebei Medical University, Shijiazhuang, Hebei 050000 China.

[†] Min Xian and Yingtao Zhang are co-first authors.

* To whom correspondence should be addressed (hengda.cheng@usu.edu).

Article	Type	Year	Category	# of images/Availability	Metrics
Xian, et al. [3]	F	2015	Graph-based	184/private	TP, FP, SI, HD, MD
Shao, et al. [18]	F	2015	Graph-based	450/private	TP, FP, SI
Huang, et al.[4]	S	2014	Graph-based	20/private	ARE,TPVF, FPVF,FNVF
Pons, et al [5]	S	2014	Deformable models	163/private	Sensitivity, ROC area
Xian, et al. [6]	F	2014	Graph-based	131/private	SI, FP, AHE
Kuo, et al.[7]	S	2014	Deformable models	98/private	DS
Torbati, et al.[8]	S	2014	Neural network	30/private	JI
Moon, et al. [9]	S	2014	Fuzzy C-means	148/private	Sensitivity and FP
Jiang, et al.[10]	S	2012	Adaboost	112/private	Mean overlap ratio
Shan, et al. [11]	F	2012	Neural network	60/private	TP, FP, FN, HD, MD
Yang, et al.[12]	S	2012	Naive Bayes classifier	33/private	FP
Shan, et al.[13]	F	2012	Neutrosophic L-mean	122/private	TP, FP, FN, SI, HD, and MD
Liu, et al. [14]	S	2012	Cellular automata	205/private	TP, FP,FN, SI
Hao, et al. [57]	F	2012	DPM + CRF	480/private	JI
Gao, et al. [15]	S	2012	Normalized cut	100/private	TP, FP, SI, HD, MD
Hao, et al. [38]	F	2012	Hierarchical SVM + CRF	261/private	JI
Liu, et al. [16]	S	2010	Level set-based	79/private	TP, FP, SI
Gómez, et al.[17]	S	2010	Watershed	50/private	Overlap ratio, NRV and PD

Table 1. Recently published approaches. F: fully automatic, S: semi-automatic; SVM: support vector machine, CRF: conditional random field, DPM: deformable part model, TP: true positive, FP: false positive, SI: similarity, HD: Hausdorff distance, MD: mean distance, DS: Dice similarity, JI: Jaccard Index, ROC: Receiver operating characteristic, ARE: average radial error, TPVF : true positive volume fraction , FPVF : false positive volume fraction, FNVF: false negative volume fraction, PR: precision ratio, MR: match rate, NRV: normalized residual value, and PD: proportional distance.

popular quantitative metrics.

We also put the BUS dataset and the performances of the five approaches on http://cvprip.cs.usu.edu/busbench. To the authors' best knowledge, this is the first attempt to benchmarking the BUS image segmentation methods. With the help of this benchmark, researchers can compare their methods with other algorithms, and find the primary and essential factors improving the segmentation performance.

The paper is organized as follows: in section II, a brief review of BUS image segmentation approaches is given; in section III, the set-up of the benchmark is presented; in section IV, the experiments and discussions are conducted; and in section V, the conclusion is presented.

II. RELATED WORK

Many BUS segmentation approaches have been studied in the last two decades, and have been proved to be effective on their private datasets. In this section, we present a brief review of automatic BUS image segmentation approaches. For more details about BUS image segmentation approaches, refer the survey paper [19].

We classify the BUS image segmentation approaches in four categories: (1) deformable models, (2) graph-based approaches, (3) learning-based approaches, and (4) classical approaches (e.g., thresholding, region growing, and watershed).

Deformable models (DMs): According to the ways of representing the curves and surfaces, we can generally classify DMs into two subcategories: (1) the parametric DMs (PDMs) and (2) the geometric DMs (GDMs).

In PDMs-based BUS image segmentation approaches, the main work was focused on generating good initial tumor boundary. Madabhushi et al. [20] proposed a fully automatic approach for BUS tumor segmentation by initializing PDMs using the boundary points produced in tumor localization step; and the balloon forces were employed in the extern forces. Chang et al.

[21] utilized the sticks filter [22] to enhance edge and reduce speckle noise before using the PDMs. Huang et al. [23] proposed an automatic BUS image segmentation approach by using the gradient vector flow (GVF) model [24], and the initial boundary was obtained by using the watershed approach.

In GDMs-based BUS image segmentation approaches, many methods focused on dealing with the weak boundary and inhomogeneity of BUS images. Gomez et al. [25] proposed a BUS image segmentation approach based on the active contour without edges (ACWE) model [26] which defined the stopping term on Mumford-Shah technique. The initial contour was a five-pixel radius circle centered at a point in the tumor marked by the user. Daoud et al. [27] built a two-fold termination criterion based on the signal-to-noise ratio and local intensity value. Gao et al. [28] proposed a level set approach based on the method in [29] by redefining the edge-based stop function using phase congruency [30] which was invariant to intensity magnitude, and integrated the GVF model into the level set framework. Liu et al. [16] proposed a GDMs-based approach which enforced priors of intensity distribution by calculating the probability density difference between the observed intensity distributions and the estimated Rayleigh distribution.

Graph-based approaches: graph-based approaches gain popularity in BUS image segmentation because of their flexibility and efficient energy-optimization. The Markov random field - Maximum a posteriori - Iterated Conditional Mode (**MRF-MAP-ICM**) and the **Graph cuts** are the two major frameworks in graph-based approaches.

MRF-MAP-ICM: Boukerroui et al. [31] stated that healthy and pathological breast tissues presented different textures on BUS images, and proposed an improved method in [32] by modeling both intensity and texture distributions in the likelihood energy; they also assumed that the texture features represented by using co-occurrence matrix follow the Gaussian distribution; and the parameters were estimated in a way similar

to the one in [32]. In [33], the Gaussian parameters in the likelihood energy were defined globally and specified manually. [34] proposed a one-click user interaction to estimate Gaussian parameters automatically.

Graph cuts: Xian et al. [3] proposed a fully automatic BUS image segmentation framework in which the graph cuts energy modeled the information from both the frequency and space domains. The data term (likelihood energy) modeled the tumor pose, position and intensity distribution. [35] built the graph on image regions, and initialized it by specifying a group of tumor regions (F) and a group of background regions (B). The weight of any *t-link* was set to ∞ if the node belonged to $F \cap B$, and all the other weights of *t-links* were set to 0; and the region intensity difference and edge strength discussed in [36] were applied to define the weight function of the smoothness term (prior energy). [35] proposed a discriminative graph cut approach in which the data term was determined online by a pre-trained Probabilistic Boosting Tree (PBT) [37] classifier.

In [38], hierarchical multiscale superpixel classification framework was proposed to define the data term. The hierarchical classifier had four layers (20, 50, 200, and 800 superpixels/nodes) built by using the normalized cut and k-means for multiscale representation; the histogram difference (Euclidean distance) between adjacent superpixels was used to define the weights in the smoothness term.

Low optimization speed and locally optima are the two main drawbacks of the MRF-MAP-ICM; while the "shrinking" problem is the main disadvantage of Graph cuts-based approaches.

Learning based approaches: both supervised and unsupervised learning approaches have been applied to solve the BUS image segmentation problem. Unsupervised approaches are simple and fast, and commonly utilized as preprocessing to generate candidate image regions. Supervised approaches are good in integrating features at different levels, but not good in applying boundary constraints to generate accurate tumor boundary.

Clustering: Xu et al. [39] proposed a BUS image segmentation method applying the spatial FCM (sFCM) [40] to the local texture and intensity features. In sFCM, the membership value of each point was updated by using its neighbors' membership values. In [39], the number of clusters was set as 2, and the membership values were assigned by using the modes of image histogram as the initial cluster centers. In [41], FCM was applied to pixel intensities for generating image regions in four clusters; then morphology, location and size features were measured for each region; a linear regression model trained on the features was employed to produce the tumor likelihoods for all regions, and the region with the highest likelihood was considered as a tumor. Moon et al. [9] applied FCM to image regions produced by using the mean shift method; the number of clusters was set to 4, and the regions belonging to the darkest cluster were extracted as the tumor candidates. Shan et al. [13] extended the FCM and proposed the neutrosophic l-means (NLM) clustering to deal with the weak boundary problem in BUS image segmentation; and it took the indeterminacy of membership into consideration.

SVM and NN: Liu et al. [42] trained a SVM classifier using local image features to classify small image lattices (16×16) into the tumor or non-tumor classes; the radius basis function (RBF) was utilized; and 18 features (16 features from co-occurrence matrix and the mean and variance of the intensities) were extracted from a lattice. Jiang et al. [10] trained Adaboost classifier using 24 Haar-like features [43] to generate a set of candidate tumor regions and trained SVM to determine the false positive and true positive regions. [44] proposed an NN-based method to segment 3D BUS images by processing 2D images slices using local image features. Othman et al. [45] trained two ANNs to determine the best-possible threshold. The ANN had 3 layers, 60 nodes in hidden layer, one node in the output layer. The first ANN used the Scale Invariant Feature Transform (SIFT) descriptors as the inputs; and the second employed the texture features from the Grey Level Co-occurrence Matrix (GLCM) as the inputs. [11] trained an ANN to conduct pixel-level classification by using the joint probability of intensity and texture [20] and two new features: the phase in the max-energy orientation (PMO) and radial distance (RD). The ANN had 6 hidden nodes and 1 output node.

Deep Learning-based approaches have been reported to achieve state-of-the-art performance for many medical tasks such as prostate segmentation [46], cell tracking [47], muscle perimysium segmentation [48], brain tissue segmentation [49], breast tumor diagnosis [50], etc. Deep learning models have great potential to achieve good performance because of their ability to characterize big image variations and to learn compact image representation using sufficiently large BUS image dataset. Deep learning architectures based on convolutional neural networks (CNNs) and recurrent neural networks (RNNs) are employed in medical image segmentation [46 - 50].

Classical approaches: Three most popular classical approaches were applied to BUS image segmentation: thresholding, region growing and watershed.

In BUS image segmentation, thresholding was often used as a pre-processing step for tumor localization. Region growing extracts image regions by starting from a set of pixels (called *seeds*) and growing seeds to large regions based on predefined *growth criteria*. In [11], Shan et al. proposed an automatic seed generation approach. In [52], Kwak et al. defined the cost of growing a region by modelling common contour smoothness and region similarity (mean intensity and size).

Watershed could produce more stable results than thresholding and region growing approaches, and selecting the marker(s) is the key issue in watershed segmentation. Huang et al. [40] selected the markers based on grey level and connectivity. [54] applied watershed to determine the boundaries on binary image. The markers were set as the connected dark regions on the binary image. [56] applied watershed and post-refinement based on grey level and location to generate candidate tumor regions.

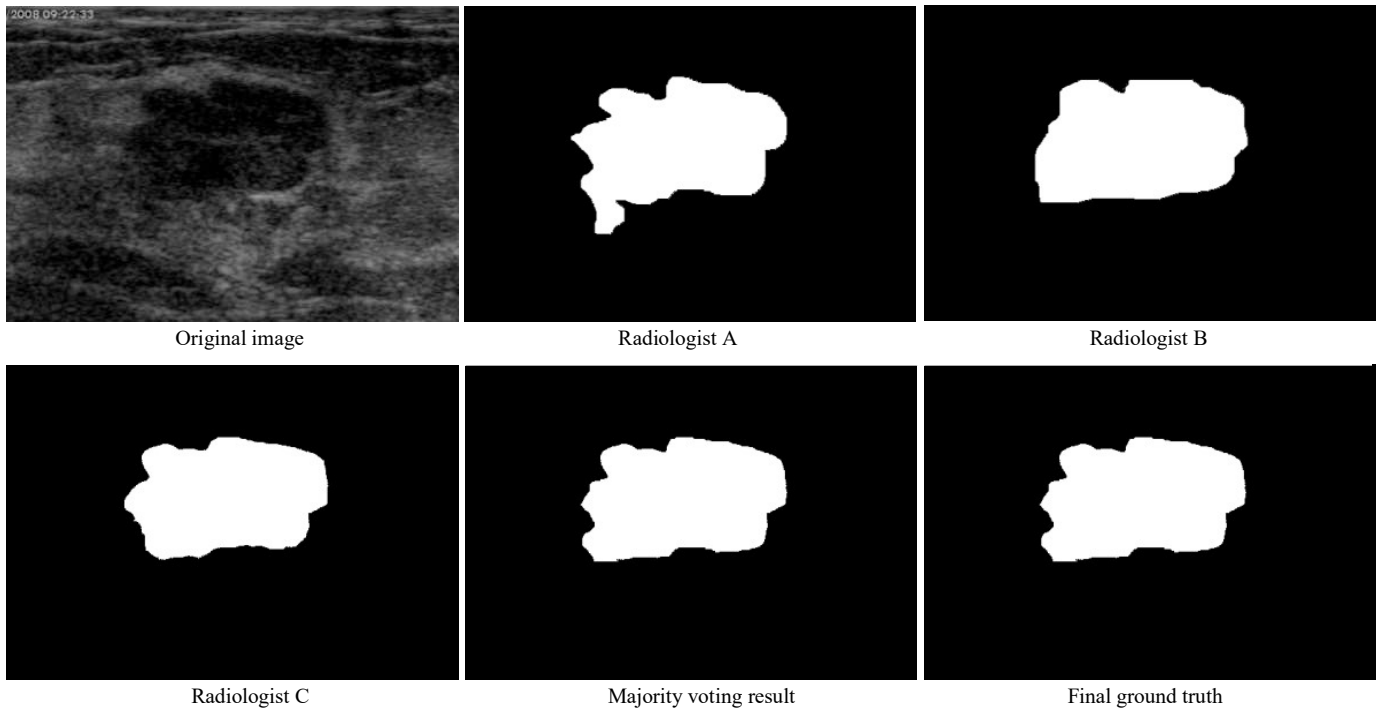In Table 1, we list the brief information of 18 approaches published recently.

Original image	Radiologist A	Radiologist B
Radiologist C	Majority voting result	Final ground truth

Figure 1. Ground truth generation

III. BENCHMARK SETUP

A. Approaches and Setup

We obtained permissions from the developers of five state-of-the-art BUS segmentation methods [3, 11, 14, 16, 18] to use their codes. Approaches in [14] and [16] are interactive, and both need operator to specify regions of interest (ROIs) manually; and the other three [3, 11, 18] are fully automatic.

[16] is a level set-based segmentation approach and sets the initial tumor boundary by using user-specified ROI. The maximum number of iterations is set to 450 as the stopping criterion. [14] is based on cell competition and uses the pixels on the boundary of the ROI specified by user as the background seeds and pixels on an adaptive cross at the ROI center as the tumor seeds. [11] utilizes predefined reference point (center of the upper part of the image) for seed generation and pre-trained tumor grey-level distribution for texture feature extraction. We use the same reference point defined in [11] and the predefined grey-level distribution provided by the authors; 10-fold cross-validation is employed to evaluate the overall segmentation performance. [3] and [18] are two graph-based fully automatic approaches. In our experiments, we adopt all the parameters from the original papers correspondingly.

B. Datasets and Ground Truth Generation

Our BUS image dataset has 562 images. The images are collected by the Second Affiliated Hospital of Harbin Medical University, the Affiliated Hospital of Qingdao University, and the Second Hospital of Hebei Medical University using multiple ultrasound devices: GE VIVID 7 and LOGIQ E9, Hitachi EUB-6500, Philips iU22, and Siemens ACUSON S2000. The images from different resources may be valuable for testing the robustness of the algorithms. Informed consents to the protocol from all patients were acquired. The privacy of the patients is well protected.

Four experienced radiologists are involved in the ground truth generation; three radiologists read and delineated each tumor boundary individually, and the fourth one (senior expert) will judge if the majority voting results need adjustment. The complete procedures of the ground truth generation are as follows.

Step 1: every of three experienced radiologists delineates each tumor boundary manually, and three delineation results will be produced for each BUS image.

Step 2: view all pixels inside/on the boundary as tumor region, and outside pixels as background; conduct majority voting to generate the preliminary result for each BUS image.

Step 3: a senior expert will read each BUS image and refer its corresponding preliminary result to decide if it needs any adjustment.

Step 4: label tumor pixel as **1** and background pixel as **0**; and generate a binary and uncompressed image to save the ground truth for each BUS image.

An example of the ground truth generation is in Figure 1.

C. Quantitative Metrics

Among the five approaches, two of them [14, 16] are semi-automatic and user predefined ROI needs to be set before the segmentation; while the other three approaches [3, 11, 18] are fully automatic. The performance of semi-automatic approaches may vary with different user interactions. It is difficult and meaningless to compare semi-automatic methods with fully
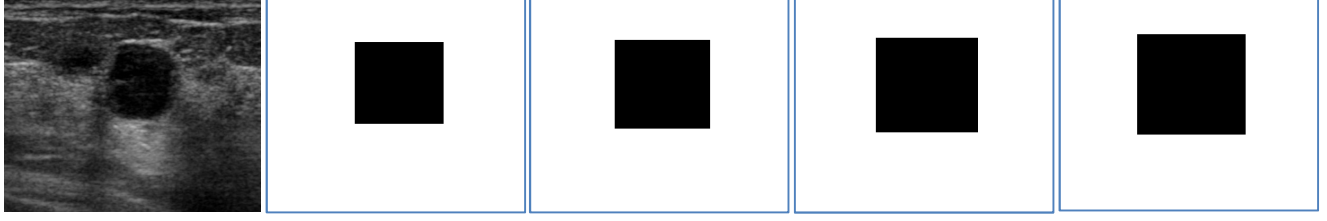
Figure 2. An example of 4 ROIs with *LRs* of 1.3, 1.5, 1.7 and 1.9, respectively.

automatic methods; therefore, we will compare the methods in two categories separately.

In the evaluation of semi-automatic approaches, we compare the segmentation performances of the two methods using the same set of ROIs, and evaluate the sensitivity of the methods to ROIs with different looseness ratio (*LR*) defined by

$$LR = \frac{BD}{BD_0}$$

where BD_0 is the size of the bounding box of the ground truth and is used as the baseline, and BD is the size of a ROI containing BD_0. We produce 10 groups of ROIs with different *LRs* automatically using the approach described in [55]: move the four sides of a ROI toward the image borders to increase the looseness ratio; and the amount of the move is proportional to the margin between the side and the image border. The *LR* of the first group is 1.1; and the *LR* of each of the other groups is 20% larger than that of its previous group. A BUS image and its four ROIs with different *LRs* are shown in Figure 2.

The method in [11] is fully automatic, involves neural network training and testing, and a 10-fold cross validation strategy is utilized to evaluate its performance. Methods in [3, 18] need no training and operator interaction. All experiments are performed using a windows-based PC equipped with a dual-core (2.6 GHz) processor and 8 GB memory. The performances of these methods are validated by comparing the results with the ground truths.

Both area and boundary error metrics are employed to assess the performance of the five approaches. The area error metrics include the true positive ratio (TPR), false positive ratio (FPR), Jaccard index (JI), Dice's coefficient (DSC), and area error ratio (AER)

$$TPR = \frac{|A_m \cap A_r|}{|A_m|} \qquad (1)$$

$$FPR = \frac{|A_m \cup A_r - A_m|}{|A_m|} \qquad (2)$$

$$JI = \frac{|A_m \cap A_r|}{|A_m \cup A_r|} \qquad (3)$$

$$DSC = \frac{2|A_m \cap A_r|}{|A_m| + |A_r|} \qquad (4)$$

$$AER = \frac{|A_m \cup A_r| - |A_m \cap A_r|}{|A_m|} \qquad (5)$$

where A_m is the pixel set of the tumor region of the ground truth,

A_r is the pixel set of the tumor region generated by a segmentation method, and $|\cdot|$ indicates the number of elements of a set. TPR, FPR and AER take values in [0, 1]; and FPR could be greater than 1 and takes value in [0, $+\infty$). Furthermore, Hausdorf error (HE) and mean absolute error (MAE) are used to measure the worst possible disagreement and the average agreement between two boundaries, respectively. Let C_m and C_r be the boundaries of tumor in the ground truth and the segmentation result, respectively. The HE is defined by

$$HE(C_m, C_r) = \max\{\max_{x \in C_m}\{d(x, C_r)\}, \max_{y \in C_r}\{d(y, C_m)\}\} \quad (6)$$

where *x* and *y* are the points on boundaries C_m and C_r, respectively; $d(\cdot, C)$ is the distance between a point and a boundary C as

$$d(z, C) = \min_{k \in C}\{\|z - k\|\}$$

where $\|z - k\|$ is the Euclidean distance between points *z* and *k*; and $d(z, C)$ is the minimum distance between point *z* and all points on *C*.

MAE is defined by

$$MAE(C_m, C_r) = 1/2 \left(\sum_{x \in C_m} \frac{d(x, C_r)}{n_r} + \sum_{y \in C_r} \frac{d(y, C_m)}{n_m} \right). (7)$$

In Eq. (7), n_r and n_m are the numbers of points on boundaries C_r and C_m, respectively.

The seven metrics above were discussed in [19]. For the first two metrics (TPR and FPR), each of them only measures a certain aspect of the segmentation result, and is not suitable for describing the overall performance; e.g., a high TPR value indicates that most portion of the tumor region is in the segmentation result; however, it cannot claim an accurate segmentation because it does not measure the ratio of correctly segmented non-tumor regions. The other five metrics (JI, DSC, AER, HE and MAE) are more comprehensive and effective to measure the overall performance of segmentation approaches, and are commonly applied to tune the parameters of segmentation models [3], e.g., large JI and DSC and small AER, HE and MAE values indicate the high overall segmentation performance.

Although JI, DSC, AER, HE and MAE are comprehensive metrics, we still recommend using both TPR and FPR for evaluating BUS image segmentation; since with these two metrics, we can discover some hidden characteristics that cannot be found through the comprehensive metrics. Suppose that the algorithm has low overall performance (small JI and DSC, and large AER, HE and MAE); if FPR and TPR are large, we can
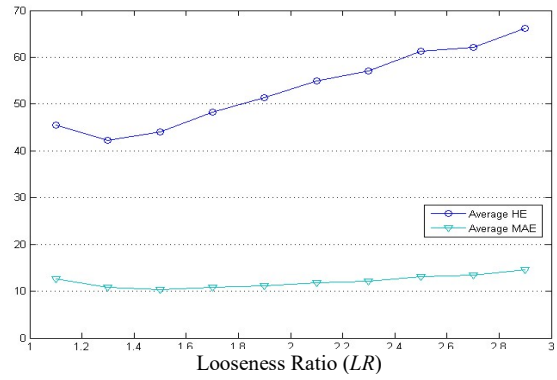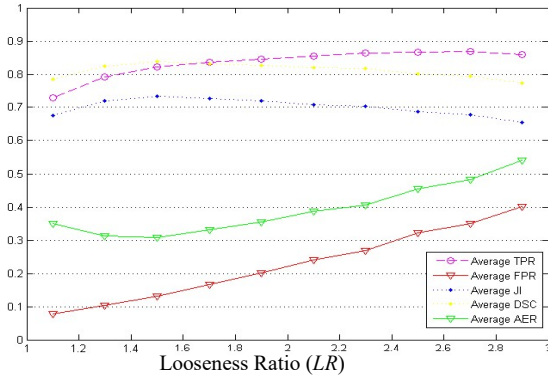
Figure 3. Segmetation results of [16]. (a) average TPR, FPR, JI, DSC and AER on different ROI sets; (b) average HE and MAE.
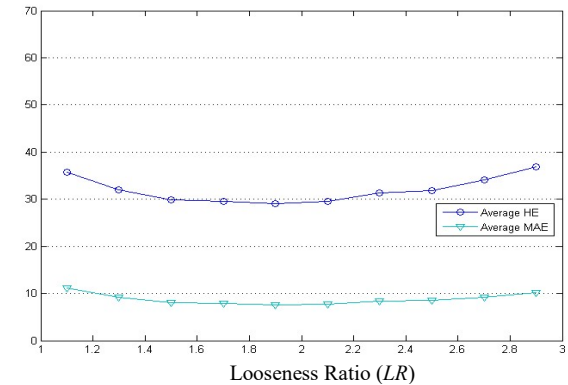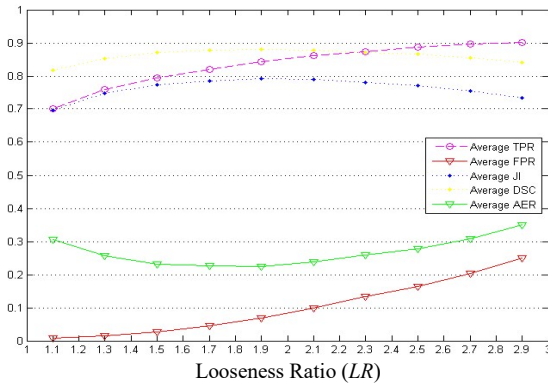

Figure 4. Segmentation results of [14]. (a) average TPR, FPR, JI, DSC and AER on different ROI sets; (b) average HE and MAE.

conclude that the algorithm has overestimated the tumor region; if both FPR and TPR are small, the algorithm has underestimated the tumor regions. The findings from TPR and FPR can guide the improvement of the BUS segmentation algorithms.

IV. APPROACHES COMPARISON AND DISCUSSIONS

In this section, we evaluate five state-of-the-art approaches [11, 14, 16, 18, 26]. For fully automatic approaches, we compare their average performances by using the seven metrics discussed in section III-C; while for semi-automatic approaches, we also evaluate their sensitivities to different *LR*s.

A. Semi-automatic Segmentation

Ten ROIs are generated automatically for each BUS image, and *LR*s range from 1.1 to 2.9 (step is 0.2). Totally, 5620 ROIs are generated for the entire BUS dataset, and we run each of the semi-automatic segmentation approach 5620 times to produce the results. All the segmentation results on the ROIs with the same *LR* are utilized to calculate the average TRP, FPR, DSC, AER, HE and MAE, respectively; and the results of the approaches in [14] and [16] are shown in Figure 3 and Figure 4, respectively.

All the average TPR and DSC values of the method in [16] are above 0.7, and its average JI values vary in the range [0.65, 0.75]. The average TPR values increase with the increasing *LR* values of ROIs. Both the average JI and DSC values tend to increase firstly, and then decrease with the increasing *LR*s of ROIs. FPR, AER and HE have low average values when the *LR*s are small, which indicate that the high performance of

method in [16] can be achieved by using tight ROIs; however, the values of the three metrics increase almost linearly with the *LR*s of ROIs when the looseness is greater than 1.3; this observation shows that the overall performance of [16] drops rapidly by using large ROIs above a certain level of *LR*. The average MAE values decrease firstly, and then increase and vary with the *LR*s in a small range. Four metrics (average JI, DSC, AER and MAE) reach their optimal values at the *LR* of 1.5 (Table 2).

The segmentation results of the approach in [14] are demonstrated in Figure 4. All average JI values are between 0.7 and 0.8; and all average DSC values are between 0.8 and 0.9. As the method in [16], the average TPR values of the method in [14] are above 0.7; and increase with *LR*s of ROIs; the average JI and DSC values increase firstly, and then decrease; the average FPR values increase with the increasing looseness of ROIs; and the average DSC, HE and MAE decrease firstly, and then increase. Five metrics (average JI, DSC, AER, HE and MAE) reach their optimal values at the *LR* of 1.9 (Table 2).

As shown in Figures 3 and 4 and Table 2, the two approaches achieve their best performances with different *LR*s (1.5 and 1.9 respectively). We can also observe the following facts:

- [16] and [14] are quite sensitive to the sizes of ROIs; their performances vary greatly at different *LR*s.
- Every approach achieves the best performance at a certain value of *LR*; however, not at the lowest looseness level (1.1).

Methods	Looseness Ratio	Area error metrics					Boundary error metrics		Time
		Ave. TPR	Ave. FPR	Ave. JI	Ave. DSC	Ave. AER	Ave. HE	Ave. MAE	Ave. Time (s)
[16]	1.1	0.73	**0.08**	0.67	0.78	0.35	45.4	12.6	18
	1.3	0.79	0.10	0.72	0.82	0.31	**42.2**	10.9	22
	1.5	0.82	0.13	**0.73**	**0.84**	**0.31**	44.0	**10.4**	27
	1.7	0.83	0.17	**0.73**	0.83	0.33	48.3	10.9	27
	1.9	0.85	0.20	0.72	0.83	0.36	51.3	11.2	30
	2.1	0.86	0.24	0.71	0.82	0.39	54.9	11.7	30
	2.3	0.86	0.27	0.70	0.82	0.41	57.0	12.1	36
	2.5	**0.87**	0.32	0.69	0.80	0.46	61.3	13.1	39
	2.7	**0.87**	0.35	0.68	0.79	0.48	62.1	13.4	40
	2.9	0.86	0.40	0.66	0.77	0.54	66.2	14.6	44
[14]	1.1	0.70	**0.01**	0.70	0.82	0.31	35.8	11.1	487
	1.3	0.76	0.02	0.75	0.85	0.26	32.0	9.1	467
	1.5	0.79	0.03	0.77	0.87	**0.23**	29.9	8.1	351
	1.7	0.82	0.05	**0.79**	**0.88**	**0.23**	29.5	7.8	341
	1.9	0.84	0.07	**0.79**	**0.88**	**0.23**	**29.0**	**7.6**	336
	2.1	0.86	0.10	**0.79**	**0.88**	0.24	29.5	7.7	371
	2.3	0.87	0.13	0.78	0.87	0.26	31.3	8.3	343
	2.5	0.89	0.16	0.77	0.87	0.28	31.9	8.5	365
	2.7	**0.90**	0.20	0.75	0.85	0.31	34.1	9.2	343
	2.9	**0.90**	0.25	0.73	0.84	0.35	36.9	10.2	388

Table 2. Quantitative results of [14] and [16] using 10 *LR*s of ROI.

Methods	Area error metrics					Boundary error metrics		Time
	Ave. TPR	Ave. FPR	Ave. JI	Ave. DSC	Ave. AER	Ave. HE	Ave. MAE	Ave. Time (s)
Fully automatic approaches								
[3]	**0.81**	**0.16**	**0.72**	**0.83**	**0.36**	**49.2**	**12.7**	7
[11]	**0.81**	1.06	0.60	0.70	1.25	107.6	26.6	16
[18]	0.67	0.18	0.61	0.71	0.51	69.2	21.3	20
Semi-automatic approaches								
[16], *LR* = 1.5	0.82	0.13	0.73	0.84	0.31	44.0	10.4	27
[24], *LR* = 1.9	**0.84**	**0.07**	**0.79**	**0.88**	**0.23**	**29.0**	**7.6**	371

Table 3. Results of three fully automatic tumor segmentation approaches [3, 11, 18] and the average optimal performances of [14] and [16].

- The performances of the two approaches drop if the looseness level is greater than a certain value; and the performance of the method [14] drops much slower than that of the method in [11].
- Set 1.9 as the optimal *LR* for [14] and 1.5 for [16]; and [14] achieves better optimal average performance than that of [16].
- The running time of the approach in [16] is proportional to the size of a user specified ROI, while there is no such relationship of the running time of the approach in [14].
- The running time of the approach in [14] is slower than that of the approach in [16] by one order of the magnitude.

B. Fully Automatic Segmentation

[3], [11] and [18] are three fully automatic approaches and their segmentation performances are shown in Table 3. The method in [3] achieves better performance than that of the methods in [11] and [18] on all five comprehensive metrics and its average FPR is also the lowest. The method in [11] has the same average TPR value as the method in [3]; however, its average FPR value reaches 106.5% which is almost six times larger than that of the method in [3]; the high average FPR and AER values of the method in [11] indicate that a large portion of non-tumor regions are misclassified as tumor regions. Among the three fully automatic approach, [18] has the lowest average TPR (62.1%); however, it outperforms the method in [11] on all the other six metrics.

In Table 3, we also show the average optimal performances of the methods in [16] and [14] at the *LR*s of 1.5 and 1.9, respectively. Their performances outperform that of all the fully automatic approaches, and [14] achieves the best performance among them; but the two approaches are much slower than the fully automatic approaches, and operators are impossible to know what the best ROI sizes are beforehand which can lead to the best segmentation results.

C. Discussions

As discussed in [19], many semi-automatic segmentation approaches are utilized for BUS image segmentation. User interactions (setting seeds and/or ROIs) are required in these approaches, and could be useful for segmenting BUS images with extremely low quality. As shown in Table 3, the two interactive approaches could achieve better performances than many fully automatic approaches if the ROI size is set properly.

Figures 3 and 4 also demonstrate that the two approaches achieve different performances using different sizes of ROIs. Therefore, the major issue in semi-automatic approaches is how to determine the best ROIs/seeds. But such issue has been neglected before; most semi-automatic approaches were only focused on improving segmentation performance by designing

more complex features and segmentation models, and did not consider user interaction as an important factor that could affect the segmentation performance. Hence, we recommend researchers that they should consider this issue when they develop semi-automatic approaches. Two possible solutions could be applied to solve this issue. First, for a given approach, we could choose the best *LR* by running experiments on a given BUS image training set (like section IV-A) and apply the *LR* to the test set. Second, like the interactive segmentation approach in [55], we could bypass this issue by designing segmentation models less sensitive to user interactions.

Fully automatic segmentation approaches have many good properties such as operator-independence and reproducibility. The key strategy that shared by many successful fully automatic approaches is to localize the tumor ROI by modeling the domain knowledge. [11] localizes tumor ROI by formulizing the empirical tumor location, appearance and size; [22] generates tumor ROI by finding adaptive reference position; and in [18], the ROI is generated to detect the mammary layer of BUS image, and the segmentation algorithm only detects the tumor in this layer. However, in many fully automatic approaches, some inflexible constraints are utilized which lower their robustness; e.g., [11] utilizes the fixed reference position to rank the candidate regions in the ROI localization process, and achieves good performance on the private BUS dataset; however, it fails to localize 8 images of the benchmark. As discussed in [19], to avoid the degradation of segmentation performance for BUS images not collected under the same controlled settings, it is important to develop unconstrained segmentation techniques that are invariant to image settings.

As shown in Table 1, many different quantitative metrics exist for evaluating the performances of BUS image segmentation approaches. In this paper, we have applied seven metrics recommended in [19] to evaluate BUS image segmentation approaches. As shown in Figures 3 and 4, average JI, DSC and AER have the same trend, and each of them is sufficient to evaluate the area error comprehensively.

V. CONCLUSION

In this paper, we establish a BUS image benchmark and present the comparing results of five state-of-the-art BUS segmentation approaches; two of them are semi-automatic and others are fully automatic. The BUS dataset contains 562 BUS images collected using three different ultrasound machines; therefore, the images have large variance in terms of image contrast, brightness and degree of noise, and can be valuable for testing the robustness of the algorithms as well. In the five approaches, two of them [3, 18] are graph-based approaches, [11] is ANN-based approach, [16] is a level set-based segmentation approach, and [14] is based on cell competition.

The quantitative analysis of the considered approaches highlights the following important issues.

- By using the benchmark, no approaches in this study can achieve the same performances reported in their original papers.
- The two semi-automatic approaches are quite sensitive to user interaction (*LR*).

- The approach modelling knowledge that is more robust can achieve better performance on image dataset with large variance in image quality.
- The approach based on strong constraints such as predefined reference point, intensity distribution, and ROI size cannot handle BUS images from different sources well.
- The quantitative metrics such as JI, DSC, AER, HE and MAE are more comprehensive and effective to measure the overall segmentation performance than TPR and FPR; however, TPR and FPR are also useful for developing and improving algorithms.

In addition, the benchmark should be and will be expanded continuously.

VI. ACKNOWLEDGEMENT

REFERENCES

[1] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2015," CA-Cancer J. Clin., vol. 65, no. 1, pp. 5-29, 2015.

[2] H. D. Cheng, J. Shan, W. Ju et al., "Automated breast cancer detection and classification using ultrasound images: A survey," Pattern Recognt., vol. 43, no. 1, pp. 299-317, Jan, 2010.

[3] M. Xian, Y. Zhang, and H. D. Cheng, "Fully automatic segmentation of breast ultrasound images based on breast characteristics in space and frequency domains," Pattern Recognit., vol. 48, no. 2, pp. 485-497, 2015.

[4] Q. Huang, X. Bai, Y. Li et al., "Optimized graph-based segmentation for ultrasound images," Neurocomputing, vol. 129, pp. 216-224, 2014.

[5] G. Pons, R. Martí, S. Ganau et al., "Computerized Detection of Breast Lesions Using Deformable Part Models in Ultrasound Images," Ultrasound in medicine & biology, 2014.

[6] M. Xian, H. Cheng, and Y. Zhang, "A Fully Automatic Breast Ultrasound Image Segmentation Approach Based on Neutro-Connectedness." In ICPR, 2014, pp. 2495-2500.

[7] H.-C. Kuo, M. L. Giger, I. Reiser et al., "Segmentation of breast masses on dedicated breast computed tomography and three-dimensional breast ultrasound images," Journal of Medical Imaging, vol. 1, no. 1, pp. 014501(1-12), 2014.

[8] N. Torbati, A. Ayatollahi, and A. Kermani, "An efficient neural network based method for medical image segmentation," Computers in Biology and Medicine, vol. 44, pp. 76-87, 2014.

[9] W. K. Moon, C.-M. Lo, R.-T. Chen et al., "Tumor detection in automated breast ultrasound images using quantitative tissue clustering," Medical Physics, vol. 41, no. 4, pp. 042901, 2014.

[10] P. Jiang, J. Peng, G. Zhang et al., "Learning-based automatic breast tumor detection and segmentation in ultrasound images." in ISBI, 2012, pp. 1587-1590.

[11] J. Shan, H. D. Cheng, and Y. X. Wang, "Completely Automated Segmentation Approach for Breast Ultrasound Images Using Multiple-Domain Features," Ultrasound Med. Biol. , vol. 38, no. 2, pp. 262-275, Feb, 2012.

[12] M.-C. Yang, C.-S. Huang, J.-H. Chen et al., "Whole breast lesion detection using naive bayes classifier for portable ultrasound," Ultrasound in medicine & biology, vol. 38, no. 11, pp. 1870-1880, 2012.

[13] J. Shan, H. Cheng, and Y. Wang, "A novel segmentation method for breast ultrasound images based on neutrosophic l-means clustering," Medical Physics, vol. 39, no. 9, pp. 5669-5682, 2012.

[14] Y. Liu, H. Cheng, J. Huang et al., "An effective approach of lesion segmentation within the breast ultrasound image based on the cellular automata principle," Journal of Digital Imaging, vol. 25, no. 5, pp. 580-590, 2012.

[15] L. Gao, W. Yang, Z. Liao et al., "Segmentation of ultrasonic breast tumors based on homogeneous patch," Medical Physics, vol. 39, no. 6, pp. 3299-3318, 2012.

[16] B. Liu, H. Cheng, J. Huang et al., "Probability density difference-based active contour for ultrasound image segmentation," Pattern Recognit., vol. 43, no. 6, pp. 2028-2042, 2010.

[17] W. Gómez, L. Leija, A. Alvarenga et al., "Computerized lesion segmentation of breast ultrasound based on marker-controlled watershed transformation," Medical Physics, vol. 37, no. 1, pp. 82-95, 2010.

[18] H. Shao, Y. Zhang, M. Xian et al., "A saliency model for automated tumor detection in breast ultrasound images.", in: IEEE ICIP, pp. 1424-1428, 2015.

[19] M. Xian, Y. Zhang, H. Cheng et al., "Automatic Breast Ultrasound Image Segmentation: A Survey," arXiv preprint arXiv:1704.01472, 2017.

[20] A. Madabhushi, and D. N. Metaxas, "Combining low-, high-level and empirical domain knowledge for automated segmentation of ultrasonic breast lesions," IEEE Trans. Med. Imaging, vol. 22, no. 2, pp. 155-169, Feb, 2003.

[21] R.-F. Chang, W.-J. Wu, W. K. Moon et al., "Segmentation of breast tumor in three-dimensional ultrasound images using three-dimensional discrete active contour model," Ultrasound in medicine & biology, vol. 29, no. 11, pp. 1571-1581, 2003.

[22] R. N. Czerwinski, D. L. Jones, and W. D. O'Brien Jr, "Detection of lines and boundaries in speckle images-application to medical ultrasound," IEEE Trans. Med. Imaging, vol. 18, no. 2, pp. 126-136, 1999.

[23] Y.-L. Huang, and D.-R. Chen, "Automatic contouring for breast tumors in 2-D sonography.", in: IEEE EMBS, 2006, pp. 3225-3228.

[24] C. Xu, and J. L. Prince, "Generalized gradient vector flow external forces for active contours," Signal Processing, vol. 71, no. 2, pp. 131-139, 1998.

[25] W. Gómez, A. Infantosi, L. Leija, W. Pereira, Active Contours without Edges Applied to Breast Lesions on Ultrasound, in: Springer MEDICON, 2010, pp. 292-295.

[26] T.F. Chan, L.A. Vese, Active contours without edges, IEEE TIP 10 (2001) 266-277.

[27] M.I. Daoud, M.M. Baba, F. Awwad, M. Al-Najjar, E.S. Tarawneh, Accurate Segmentation of Breast Tumors in Ultrasound Images Using a Custom-Made Active Contour Model and Signal-to-Noise Ratio Variations, in: IEEE SITIS, 2012, pp. 137-141.

[28] L. Gao, X. Liu, W. Chen, Phase-and gvf-based level set segmentation of ultrasonic breast tumors, J Appl. Math. 2012 (2012) 1-22.

[29] C. Li, C. Xu, C. Gui, M.D. Fox, Distance regularized level set evolution and its application to image segmentation, IEEE TIP 19 (2010) 3243-3254.

[30] P. Kovesi, Phase congruency: A low-level image invariant, Psychological Research Psychologische Forschung, 64 (2000) 136-148.

[31] D. Boukerroui, O. Basset, N. Guerin et al., "Multiresolution texture based adaptive clustering algorithm for breast lesion segmentation," EJU, vol. 8, no. 2, pp. 135-144, 1998.

[32] E. A. Ashton, and K. J. Parker, "Multiple resolution Bayesian segmentation of ultrasound images," Ultrasonic Imaging, vol. 17, no. 4, pp. 291-304, 1995.

[33] G. Xiao, M. Brady, J. A. Noble et al., "Segmentation of ultrasound B-mode images with intensity inhomogeneity correction," IEEE Trans. Med. Imaging, vol. 21, no. 1, pp. 48-57, 2002.

[34] G. Pons, J. Martí, R. Martí et al., "Simultaneous lesion segmentation and bias correction in breast ultrasound images," Pattern Recognition and Image Analysis, pp. 692-699: Springer, 2011.

[35] H.-H. Chiang, J.-Z. Cheng, P.-K. Hung et al., "Cell-based graph cut for segmentation of 2D/3D sonographic breast images," in IEEE ISBI: From Nano to Macro,2010, pp. 177-180.

[36] C.-M. Chen, Y.-H. Chou, C. S. Chen et al., "Cell-competition algorithm: A new segmentation algorithm for multiple objects with irregular boundaries in ultrasound images," Ultrasound in medicine & biology, vol. 31, no. 12, pp. 1647-1664, 2005.

[37] Z. Tu, "Probabilistic boosting-tree: Learning discriminative models for classification, recognition, and clustering," in IEEE ICCV, 2005, pp. 1589-1596.

[38] Z. Hao, Q. Wang, H. Ren et al., "Multiscale superpixel classification for tumor segmentation in breast ultrasound images," in IEEE ICIP, 2012, pp. 2817-2820.

[39] Y. Xu, "A modified spatial fuzzy clustering method based on texture analysis for ultrasound image segmentation," in IEEE ISIE, 2009, pp. 746-751.

[40] K.-S. Chuang, H.-L. Tzeng, S. Chen et al., "Fuzzy c-means clustering with spatial information for image segmentation," Computerized Medical Imaging and Graphics, vol. 30, no. 1, pp. 9-15, 2006.

[41] C. Lo, Y.-W. Shen, C.-S. Huang et al., "Computer-aided multiview tumor detection for automated whole breast ultrasound," Ultrasonic Imaging, vol. 36, no. 1, pp. 3-17, 2014.

[42] B. Liu, H. D. Cheng, J. H. Huang et al., "Fully automatic and segmentation-robust classification of breast tumors based on local texture analysis of ultrasound images," Pattern Recognit., vol. 43, no. 1, pp. 280-298, Jan, 2010.

[43] P. Viola, and M. J. Jones, "Robust real-time face detection," IJCV, vol. 57, no. 2, pp. 137-154, 2004.

[44] S. F. Huang, Y. C. Chen, and W. K. Moon, "Neural network analysis applied to tumor segmentation on 3D breast ultrasound images," in ISBI, pp. 1303-1306, 2008.

[45] A. A. Othman, and H. R. Tizhoosh, "Segmentation of Breast Ultrasound Images Using Neural Networks," Engineering Applications of Neural Networks, pp. 260-269: Springer, 2011.

[46] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in IEEE 3D Vision (3DV), 2016, pp. 565-571.

[47] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in MICCAI, 2015, pp. 234-241.

[48] Y. Xie, Z. Zhang, M. Sapkota et al., "Spatial Clockwork Recurrent Neural Network for Muscle Perimysium Segmentation," in MICCAI, 2016, pp. 185-193.

[49] W. Zhang, R. Li, H. Deng et al., "Deep convolutional neural networks for multi-modality isointense infant brain image segmentation," NeuroImage, vol. 108, pp. 214-224, 2015.

[50] J.-Z. Cheng, D. Ni, Y.-H. Chou, J. Qin, C.-M. Tiu, Y.-C. Chang, C.-S. Huang, D. Shen, C.-M. Chen, Computer-Aided diagnosis with deep learning architecture: applications to breast lesions in US images and pulmonary nodules in CT scans, Scientific Reports, 6 (2016).

[51] M. H. Yap, "A novel algorithm for initial lesion detection in ultrasound breast images," Journal of Applied Clinical Medical Physics, vol. 9, no. 4, 2008.

[52] J. I. Kwak, S. H. Kim, and N. C. Kim, "RD-based seeded region growing for extraction of breast tumor in an ultrasound volume," Computational Intelligence and Security, pp. 799-808: Springer, 2005.

[53] Y.-L. Huang, and D.-R. Chen, "Watershed segmentation for breast tumor in 2-D sonography," Ultrasound in medicine & biology, vol. 30, no. 5, pp. 625-632, May, 2004.

[54] M. Zhang, L. Zhang, and H.-D. Cheng, "Segmentation of ultrasound breast images based on a neutrosophic method," Optical Engineering, vol. 49, no. 11, pp. 117001-117001-12, 2010.

[55] M. Xian, Y. Zhang, H.-D. Cheng et al., "Neutro-connectedness cut," IEEE Transactions on Image Processing, vol. 25, no. 10, pp. 4691-4703, 2016.

[56] C. M. Lo, R. T. Chen, Y. C. Chang et al., "Multi-Dimensional Tumor Detection in Automated Whole Breast Ultrasound Using Topographic Watershed," IEEE TMI, vol. 33, no. 7, pp. 1503-1511, 2014.

[57] Z. Hao, Q. Wang, Y. K. Seong et al., "Combining CRF and multi-hypothesis detection for accurate lesion segmentation in breast sonograms," in MICCAI, 2012, pp. 504-511.