International Journal of Computer Assisted Radiology and Surgery (2020) 15:981–988
https://doi.org/10.1007/s11548-020-02158-3

ORIGINAL ARTICLE

# Two-stage ultrasound image segmentation using U-Net and test time augmentation

Mina Amiri[1] · Rupert Brooks[1,2] · Bahareh Behboodi[1] · Hassan Rivaz[1]

Received: 20 November 2019 / Accepted: 3 April 2020 / Published online: 29 April 2020


## Abstract

**Purpose** Detecting breast lesions using ultrasound imaging is an important application of computer-aided diagnosis systems. Several automatic methods have been proposed for breast lesion detection and segmentation; however, due to the ultrasound artefacts, and to the complexity of lesion shapes and locations, lesion or tumor segmentation from ultrasound breast images is still an open problem. In this paper, we propose using a lesion detection stage prior to the segmentation stage in order to improve the accuracy of the segmentation.

**Methods** We used a breast ultrasound imaging dataset which contained 163 images of the breast with either benign lesions or malignant tumors. First, we used a U-Net to detect the lesions and then used another U-Net to segment the detected region. We could show when the lesion is precisely detected, the segmentation performance substantially improves; however, if the detection stage is not precise enough, the segmentation stage also fails. Therefore, we developed a test-time augmentation technique to assess the detection stage performance.

**Results** By using the proposed two-stage approach, we could improve the average Dice score by 1.8% overall. The improvement was substantially more for images wherein the original Dice score was less than 70%, where average Dice score was improved by 14.5%.

**Conclusions** The proposed two-stage technique shows promising results for segmentation of breast US images and has a much smaller chance of failure.

**Keywords** Segmentation · Ultrasound · Detection · U-Net

## Introduction

As the most common form of cancer among women, breast cancer affects more than 8% of women worldwide and its early detection is crucial to reduce corresponding side effects and mortality rate. If the disease is detected soon enough, the patient can benefit from different treatment options. However, accurate and reliable diagnosis which is the key for an early detection should be developed to facilitate the treatment procedure. Therefore, computerized techniques with the advanced biomedical information technology have vital importance and several computer-aided diagnosis systems have been developed for breast cancer and other disorders [6].

Ultrasound (US) imaging is an effective tool for diagnosis and assessment of breast cancer. It is a low-cost imaging modality without radiation and could widely attract researchers and clinicians' attention. Studies have demonstrated that US images can discriminate benign and malignant lesions [1,18], and considerably decrease the number of unnecessary biopsies for benign lesions. Analyzing US images is a skill and experience dependent, time-consuming and subjective task. Developing automatic methods for US image processing is therefore important to aid in breast cancer detection and lesion segmentation. Many automatic algorithms have been proposed to segment the breast US images. Common conventional methods include thresholding-based [2,8], clustering-based [14,19], watershed-based [7,24] and graph-based active contour models [10,26]. Deep convolutional neural networks have also been emerged as an efficient approach for segmentation of medical images [5]; however,

✉ Mina Amiri
amirim@encs.concordia.ca

[1] Concordia University, 1493 Saint-Catherine St W, Montreal, Quebec, Canada

[2] Nuance Communications, 1500 Boulevard Robert-Bourassa, Montreal, Quebec H3A 3S7, Canada

segmentation of US images is still a challenging problem due to the speckle noise, intensity inhomogeneity and low signal to noise ratio [9].

Among all convolutional networks developed for medical image segmentation, U-Net is one of the most successful architectures [17], and many other networks have been proposed based on that. The U-Net architecture consists of three sections: the contraction, the bottleneck and the expansion parts. The main contribution of U-Net is the skip connections from contraction to expansion path. The feature maps of contraction layers are concatenated to the corresponding expansion layers. This way, the features learned while contracting the image are used to expand the feature vector to reconstruct the segmented image.

A common approach in image segmentation in general, and in breast US image segmentation in particular is a two-stage strategy: first, locating the region of interest (ROI), and then segmenting the lesion in that region. [2] for instance, located the tumor by the radial gradient index filtering technique, and then segmented the lesion using a region growing algorithm. Mathematical formulation of image features and support vector machines have been also used to automatically find the candidate regions of interests, and to generate tumor seeds [12,19].

Two-stage segmentation strategy using deep learning algorithms has been employed in computer vision applications, and several object detection methods have been proposed such as different variations of RCNN [4], YOLO [15] and SSD [13] algorithms. Training such large networks is tricky specially for the task of object detection where there are tens of different objects to detect in different locations, directions and postures. In contrast, in medical applications, the task of object detection is much simpler than that of computer vision applications, and therefore, the available object detection algorithms are too sophisticated to train and use for medical images. Scarcity of labeled training data is the main constraint for training such large and complicated networks.

In medical imaging, the two-stage approach has recently been employed to segment magnetic resonance imaging (MRI) and computed tomography (CT) images, where similar or different network architectures are used in two stages [16,20,21,25]. In these papers, the first stage could act either as a region detector or just a feature extractor.

In this paper, we propose a two-stage segmentation procedure for breast US images, where the lesions are first detected, and then segmented. U-Net architecture is used for both detection and segmentation networks, as it has been shown to train quickly and with very few images. We hypothesize that by shrinking the candidate region, the performance of the segmentation network will improve. The contributions of our paper are as follows: (1) We propose a two-stage segmentation approach for the first time for ultrasound images and show that it outperforms a single-stage U-Net by breaking the problem into two simpler problems. (2) We propose a validation method as a test time augmentation technique to further improve the results. (3) We show that training the two stages separately leads to better results than jointly optimizing the two networks.

## Materials and methods

In this paper, we propose using U-Net as a detection network before the segmentation stage. To show the impact of adding the detection stage to the segmentation workflow, we will compare the results of this two-stage strategy with the results of one-stage strategy. The schematic presentation of the two approaches is shown in Fig. 1. In the one-stage approach, the U-Net is trained using the whole images and the desired output is the annotated mask. However, in the two-stage approach, there are two networks: the first network is aimed to detect the location of the lesion, and the second network provides a pixel-level segmentation of the detected lesion.

### Dataset

We used the breast US dataset provided by [23] which includes 163 B-mode US images with benign (110) lesions or malignant (53) tumors. The mean image size is $760 \times 570$ pixels. The average size of blocks containing the lesions is 148 * 90 pixels. The smallest lesion is 35 * 17 pixels, while the largest one is 405 * 281 pixels. Figure 2 represents some example images and their associated ground truths. All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Declaration of Helsinki and its later amendments or comparable ethical standards. Informed consent was obtained from all individual participants included in the study.

### Training procedure

We used five-fold cross-validation to assess the performance of the network. The images were randomly divided into five folds. The goal of this study is to segment pixels of images into the cyst and background. The five folds were selected randomly without taking into account the size of the cysts. Four out of five folds were used for training the network, and the fifth fold was used as the test set. This procedure was repeated for five times, so that all of the data were once considered as the test set. US B-mode images and their corresponding ground truths were used to train the network. For a faster convergence, we used the weights of a pre-trained network as an initial point: We used a large two-class dataset to pre-train the network [22]. It contained 10,000 seg-

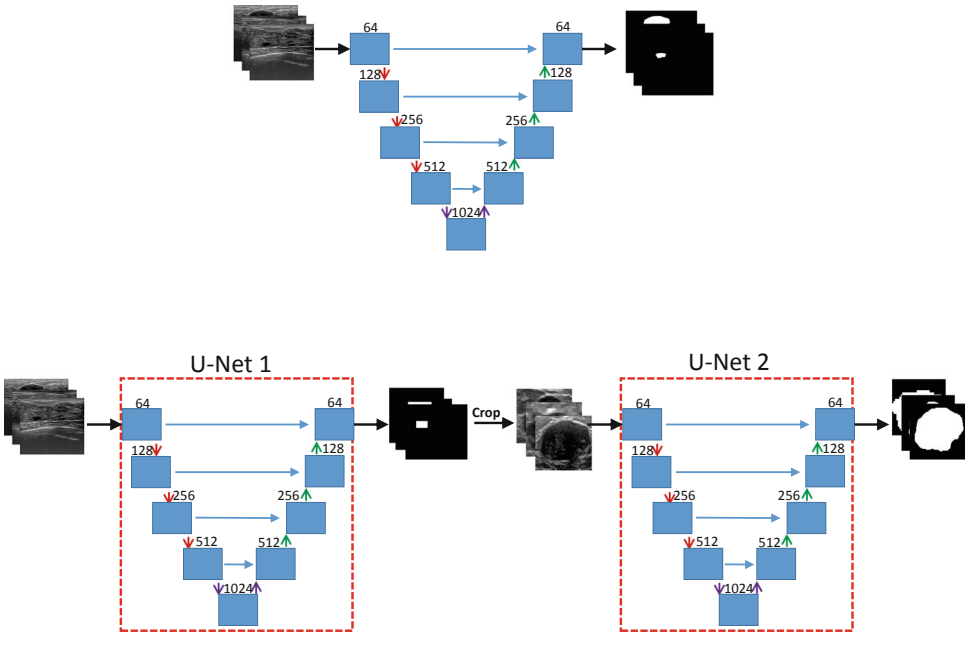


**Fig. 1** Top: The one-stage approach, using one U-Net. Bottom: The two-stage approach using one U-Net for detection and one for segmentation. Each blue box represents two convolutional layers. Red, green and blue arrows, respectively, represent maxpooling, upsampling and copy-crop-concatenating. The purple arrows show the place of dropout (50%)

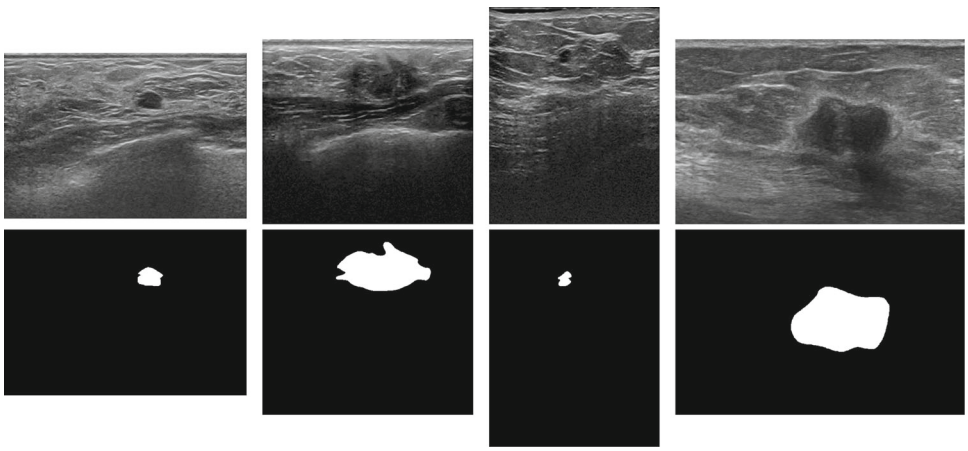

**Fig. 2** Examples from the breast US dataset used in this study, and their corresponding masks

mented images of three main categories of objects, humans and places. We selected this dataset for pre-training to have a similar segmentation problem with only 2 classes to segment. Then, the U-Net is initialized with pre-trained weights for US images. 20% of the images in the training set were considered as the validation set as a checkpoint to stop training. The stopping criteria were either 200 epochs or no improvement in validation set loss for 20 consecutive epochs.

For regularization, we used the dropout technique with the rate of 0.5 after the contraction part. We implemented on-the-fly data augmentation techniques to alleviate the problem of low training-set size. We generated smooth deformations of images using random and small degrees of shifting (with the shift range of 0.05 of the total length), rotating (random rotation in the range of 0.2 degrees), zooming (random zoom in the range of $\pm$ 0.05) and horizontal flipping. The activation function is ReLU for all layers except for the last layer which is the sigmoid activation function. The loss function is binary cross-entropy as the original U-Net paper and the network is optimized by Adam optimizer. All images were normalized to have intensity values between [0,1].

## The one-stage approach

We used the original U-Net architecture first proposed by [17], which was recently shown to outperform many more recent architectures in segmentation of US images [11]. The contraction part of the network consists of blocks of two convolutional layers with ReLU activation, followed by a maxpooling operation (Fig. 1). The network is symmetric, and there are blocks of two convolutional layers in the expansion part followed by upsampling operation. There are some skip connections that transfer the information from the contraction part to the expansion part by concatenating the learned features from the contraction part to the expansion part. The first two convolutional layers include 64 filters. In

the contraction part, the number of filters multiplies by two after each maxpooling operation, and in the expansion part, the number of filters decreases by the factor of two after each upsampling operation. The last layer is a $1 \times 1$ convolutional layer with sigmoid activation to map the feature vector to the interval of 0 and 1. For evaluation purposes, the predicted masks were binarized using the empirical threshold value of 0.5 so that pixels with the value above or equal to 0.5 were considered as 1, and pixels with the value below 0.5 were considered as 0.

### The two-stage approach

In this paper, we propose the use of the same network (U-Net) for both ROI detection and segmentation stages for breast US images. The overall architecture of our proposed framework therefore consists of two U-Net where the first network's responsibility is to detect where the lesion exists and the second network segments the detected region. To train the first network, we developed a new ground truth based on the original ground truth: instead of the actual lesion shape, we used the surrounding rectangle as the ground truth. The network was then trained the same way we trained the network in one-stage approach, e.g. applying five-fold cross-validation with the same stopping criteria, pre-trained weights and the validation set ratio.

Then, the output of the first network was considered as the input for the second network. The surrounding rectangle of detected regions was cropped and fed to the second network to be segmented. In both training and testing, if for one single image, more than one distinct regions were detected by the first network, all of the detected regions were considered as the input to the second network (Fig. 3). If no candidate region was detected by the first network, the whole image would be fed to the second network as the input image.

### Test time augmentation

The performance of the segmentation (the second network) obviously depends on how well the first network could have detected the lesion region. If the first network detects the rectangle perfectly, then the second network will be able to segment the image accurately. However, if the detection stage is not successful and the lesion is missed or part of the lesion is not detected, the second network will fail too. It is therefore necessary to find a strategy to determine whether the detection stage performs well and whether the detection results are valid.

To evaluate this, we proposed using test-time augmentation technique in two different ways. First, by augmenting the test data by shifting. If the detection is done appropriately, by shifting the image for a few pixels we expect the detected region to shift in the same direction. However, if the detection is fragile, the detected region is expected to change in an unpredictable way. We, therefore, shifted the image for ten different values ($-25, -20, -$-15, …, 15, 20, 25 pixels) and examined how the detected region changed by shifting the original image.

As the second method, we used the dropout technique at the test time. Dropout has been shown to represent the model uncertainty in deep learning [3]. By employing dropout at the test time, the network is expected to generate slightly different results at each run. If the variation between different runs for the same input is large, the result is considered uncertain. If the model returns an output with high uncertainty, one may decide to further validate it. Similar to the shifting procedure, we calculated the output of the detection network for each image in the test set, for 10 times keeping the dropout layer active at the test time. When the uncertainty between different runs is low, the detection network is considered valid. We used Dice score to measure the uncertainty between different runs. We calculated the Dice score between the output of the network when dropout layer is removed out and the output of the network with active dropout layer. When both methods declare the output as an invalid result, the performance of the first network is considered invalid.

### Evaluation

To evaluate the accuracy of the segmentation output of the two above mentioned approaches, we used Dice score which is a measure of overlap between the segmented region and the ground truth defined as:

$$\text{Dice score} = \frac{2\text{TP}}{2\text{TP} + \text{FN} + \text{FP}} \tag{1}$$

where TP is the total number of elements correctly predicted as the mask, FN is the number of elements in the ground-truth mask that are not detected by the segmentation method, and FP is the number of elements falsely detected as the mask by the method.

## Results

To ensure the fairness of the comparison between the two approaches, the same five-fold subdivision of the data was used to train and test both approaches. The average Dice score among the five folds for the one-stage approach was $79.3 \pm 3.1\%$. The average increase in the Dice score using the two-stage approach without the test-time augmentation step to validate the detection network performance was $1.2 \pm 4.5\%$.

To perform test-time augmentation, we shifted the images as explained and calculated the output of the detection net-

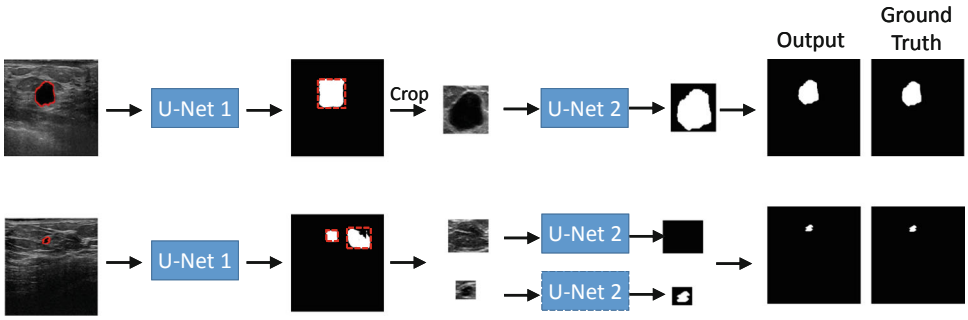


**Fig. 3** Schematic of the proposed two-stage approach, when one or more than one regions are detected in the detection stage

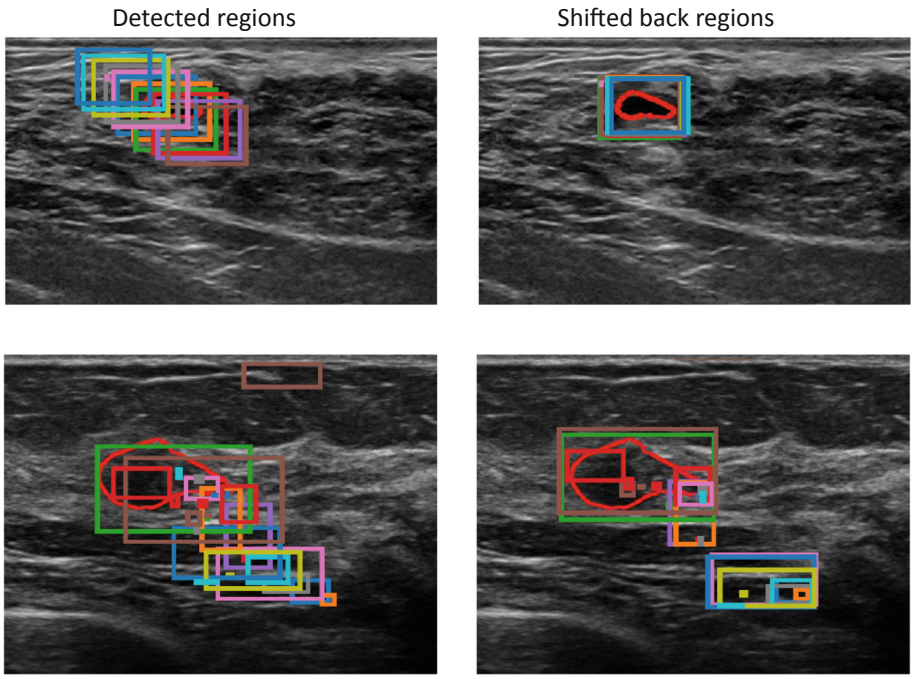


**Fig. 4** Detected regions when the original image is shifted for different number of pixels, and shifting back the detected regions. Top: An example of a presumably good detection, bottom: An example of a presumably weak detection

work for these shifted images. Figure 4 represents two example images with good and weak detection results. By shifting back the result images, we could calculate the Dice score between the detected regions and the original detected mask. The detection was then considered valid, if the average Dice score between the shifted images and the original detection was above 90%. We also used the dropout technique as the second approach to validate the performance of the detection. The detection was considered valid, if the average Dice score between the original detection and 10 different runs was above 97%. If both of two techniques confirmed invalid detection, the detection was considered invalid. The thresholds in test time augmentation techniques are tuned on the training set. By only considering images with valid detection (77% of images), the average Dice score was 86.5 ± 1.8%. The average Dice score for the same subset of images using the one-stage approach was 2.2% less than that of the two-stage approach (84.3 ± 3.6%). Table 1 represents the results of one- and two-stage approaches in different subsets of images.

When we used the original one-stage approach for images with invalid detection, the overall Dice score for all images was 81.1 ± 1.9% which is 1.8% higher than the Dice score achieved by the one-stage approach. This improvement in average Dice score was consistently seen in all five folds. It is also noteworthy that separately training the two networks yielded better results. This can be due to the fact that very deep networks are difficult to train, and this proposed solution can be viewed as a deep supervision approach.

In total, among all 163 images, the cysts were not detected by either of the approaches in five images. The cysts in 10 images were not detected using the one-stage approach, among which 5 were detected by the two-stage approach. For 3 of these detected cysts, the Dice score is noticeably high (above 80%), while for the other 2, the Dice score is below 50%. In only one image the Dice score was 0 by the two-stage approach and above 0 by the one-stage approach. However, the detection procedure was not valid in this image according to the test time evaluation method, and therefore, the two-stage approach was not considered valid for this

**Table 1** The average Dice score (%) for the two approaches in different set of images (* represents significant statistical difference, $p < 0.05$)

| | All images (no test-time augmentation) | Images with valid detection* | Images with invalid detection | Images with Dice score > 70% | Image with Dice score < 70%* |
|---|---|---|---|---|---|
| One-stage | 79.3 ± 3.1 | 84.3 ± 3.6 | 69.6 ± 6.8 | 90.2 ± 1.4 | 34.2 ± 27.0 |
| Two-stage | 80.5 ± 2.1 | 86.5 ± 1.8 | 65.2 ± 6.0 | 89.9 ± 1.2 | 48.7 ± 34.2 |

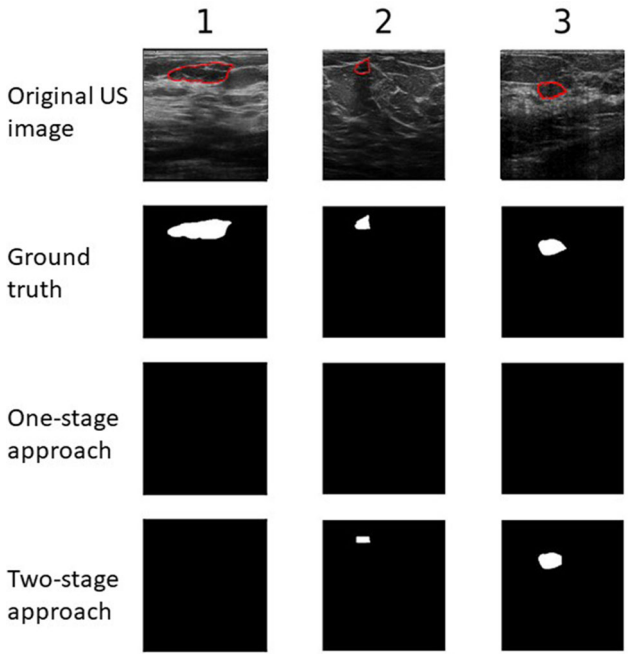


**Fig. 5** Examples of images wherein the lesion is not detected by the one-stage approach. The lesion in the first example (column) is not detected by the two-stage approach either. The lesions of the second and the third examples are detected with the proposed two-stage approach, with the Dice score of 33.9% and 86.8%, respectively

image. Figure 5 shows some examples of images for which the lesions were not detected by the one-stage approach.

We then categorized the images-based on the segmentation results of one-stage approach-into two groups of weak and strong performance. We arbitrarily considered the Dice score of less than 70% including the images with undetected lesions (Dice score = 0) as weak segmentation. In total, 32 of images had a Dice score of below 70%, with the average Dice score of 34.2%. Using the two-stage approach the average Dice score increased to 48.7%, and 16 of these images gained a Dice score of above 70%, and only three of these 32 images had worse results when applying the two-stage approach. Figure 6 shows the obtained Dice scores using one-stage and two-stage approaches, for all images with weak and strong results. As seen in Fig. 6, the Dice improvement is partly due to the better segmentation of the two-stage approach and partly due to the higher rate of detection in the images in which the cyst is missed by the one-stage approach. Interestingly, if we only consider images with a detected cyst by

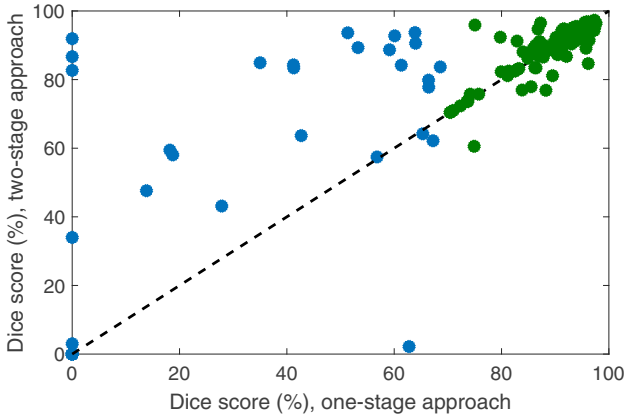


**Fig. 6** The Dice score gained using the two-stage approach when the Dice score of the one-stage approach is below 70% (blue) or above 70% (green)

one-stage approach, the average improvement in Dice score will be 22.2%, which implies the higher impact of the former factor.

For images with strong performance (Dice score above 70%), the average Dice score of the proposed two-stage approach was slightly worse than that of the one-stage approach (89.9 ± 1.2% vs. 90.2 ± 1.4%—Table 1). This could be due to slightly incorrect ROI detection in the first step (Fig. 6).

Regarding the runtime of the network, we can process images using the two-stage approach with an average speed of 6.22 fps on a TITAN V GPU. Therefore, this method can run in real-time.

## Discussion and conclusions

The performance of the one-stage approach is satisfactory in the majority of images, and as illustrated, only 32 out of 163 images had weak segmentation results. By using the proposed two-stage approach, we could substantially improve the performance for these kind of images (Fig. 6-left). For images with good segmentation results, on the other hand, the two methods work quite similarly, with the two-stage approach falling behind in a portion of images, and outperforming in some other images (Fig. 6-right). Therefore, our

proposed method particularly helps improving the segmentation in images for which the one-stage approach fails.

In breast cancer screening, it is paramount not to miss the lesion if there is any. Missing the tumor would result in not pursuing the necessary treatment procedure, which may lead to severe consequences. We could show that by using the proposed two-stage approach, 50% of the missed lesions were detected. Although, the segmentation accuracy and the Dice score were not high in all of these detected lesions, their detection per se would help in better decision for the follow-up treatment.

Although, we could not reach excellent segmentation for all images, we proposed a method to distinguish good and weak results using test-time augmentation, so that we could be sure that at least in 77% of the images, the results were trustable. By using the proposed method, we could considerably improve the Dice score.

Training a deep neural network requires a large amount of data, which is particularly difficult to access in medical applications. Higher number of images would enhance the segmentation results. In this study, we included some preprocessing steps such as detection, detection evaluation and cropping before the segmentation, and we could show that by these steps the results of segmentation would improve. We expect that by having more data, the detection network performance would improve and consequently lead to better segmentation results.

In this study, we only analyzed one breast US dataset. Including more datasets can help generalize the conclusions of this study to other anatomical structures such as heart and vascular system. Further investigation on other breast datasets as well as other anatomical regions is therefore necessary.

**Funding** This work was supported by Natural Science and Engineering Research Council of Canada (NSERC) Discovery Grant RGPIN-04136 and by Richard and Edith Strauss Foundation.

## Compliance with ethical standards

**Conflict of Interest** The authors declare that they have no conflict of interest.

## References

1. Chen CM, Chou YH, Han KC, Hung GS, Tiu CM, Chiou HJ, Chiou SY (2003) Breast lesions on sonograms: computer-aided diagnosis with nearly setting-independent features and artificial neural networks. Radiology 226(2):504–514
2. Drukker K, Giger ML, Horsch K, Kupinski MA, Vyborny CJ, Mendelson EB (2002) Computerized lesion detection on breast ultrasound. Med Phys 29(7):1438–1446
3. Gal Y, Ghahramani Z (2016) Dropout as a bayesian approximation: representing model uncertainty in deep learning. In: Proceedings of the 33rd international conference on international conference on machine learning, vol 48, pp 1050–1059
4. Girshick RB, Donahue J, Darrell T, Malik J (2013) Rich feature hierarchies for accurate object detection and semantic segmentation. In: CoRR. arXiv:1311.2524
5. Goceri E, Goceri N (2017) Deep learning in medical image analysis: recent advances and future trends. IADIS international conference big data analytics, data mining and computational intelligence 2017 (part of MCCSIS 2017). pp 305–310
6. Goceri E, Songul C (2018) Biomedical information technology: image based computer aided diagnosis systems. International conference on advanced technologies
7. Gomez W, Leija L, Alvarenga AV, Infantosi AFC, Pereira WCA (2010) Computerized lesion segmentation of breast ultrasound based on marker-controlled watershed transformation. Med Phys 37(1):82–95
8. Horsch K, Giger ML, Venta LA, Vyborny CJ (2002) Computerized diagnosis of breast lesions on ultrasound. Med Phys 29(2):157–164
9. Huang Q, Luo Y, Zhang Q (2017) Breast ultrasound image segmentation: a survey. Int J Comput Assist Radiol Surg 12(3):493–507. https://doi.org/10.1007/s11548-016-1513-1
10. Huang QH, Lee SY, Liu LZ, Lu MH, Jin LW, Li AH (2012) A robust graph-based segmentation method for breast tumors in ultrasound images. Ultrasonics 52(2):266–275
11. Leclerc S, Smistad E, Pedrosa J, Ostvik A, Cervenansky F, Espinosa F, Espeland T, Berg EAR, Jodoin P, Grenier T, Lartizien C, D'hooge J, Lovstakken L, Bernard O (2019) Deep learning for segmentation using an open large-scale dataset in 2D echocardiography. IEEE Trans Med Imaging 38(9):2198–2210
12. Liu B, Cheng H, Huang J, Tian J, Liu J, Tang X (2009) Automated segmentation of ultrasonic breast lesions using statistical texture classification and active contour based on probability distance. Ultrasound Med Biol 35(8):1309–1324. https://doi.org/10.1016/j.ultrasmedbio.2008.12.007
13. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, Berg AC (2016) SSD: single shot multibox detector. Lecture notes in computer science, pp 21–37
14. Moon WK, Lo CM, Chen RT, Shen YW, Chang JM, Huang CS, Chen JH, Hsu WW, Chang RF (2014) Tumor detection in automated breast ultrasound images using quantitative tissue clustering. Med Phys 41(4):042901
15. Redmon J, Divvala S, Girshick R, Farhadi A (2016) You only look once: unified, real-time object detection. In: IEEE conference on computer vision and pattern recognition (CVPR)
16. Reza SMS, Roy S, Park DM, Pham DL, Butman JA (2019) Cascaded convolutional neural networks for spine chordoma tumor segmentation from MRI. In: Proceedings SPIE 10953, medical imaging 2019: biomedical applications in molecular, structural, and functional imaging
17. Ronneberger O, Fischer P, Brox T (2015) U-Net: convolutional networks for biomedical image segmentation. In: International conference on medical image computing and computer assisted intervention (MICCAI)
18. Sahiner B, Chan HP, Roubidoux MA, Hadjiiski LM, Helvie MA, Paramagul C, Bailey J, Nees AV, Blane C (2010) Malignant and benign breast masses on 3D US volumetric images: effect of computer-aided diagnosis on radiologist accuracy. Radiology 242:716–724
19. Shan J, Cheng HD, Wang Y (2012) A novel segmentation method for breast ultrasound images based on neutrosophic l-means clustering. Med Phys 39(9):5669–5682
20. Shi H, Liu J, Liao H (2019) A classification and segmentation combined two-stage CNN model for automatic segmentation of brainstem. World Congr Med Phys Biomed Eng 2018:159–163

21. Wang C, MacGillivray T, Macnaught G, Yang G, Newby DE (2018) A two-stage 3D U-Net framework for multi-class segmentation on full resolution image. ArXiv arXiv:1804.04341
22. Xia C, Li J, Chen X, Zheng A, Zhang Y (2017) What is and what is not a salient object? learning salient object detector by ensembling linear exemplar regressors. In: IEEE conference on computer vision and pattern recognition (CVPR) pp 4399–4407
23. Yap MH, Pons G, Marti RM, Ganau S, Sentis M, Zwiggelaar R, Davison AK, Marti R (2018) Automated breast ultrasound lesions detection using convolutional neural networks. IEEE J Biomed Health Inf 22:1218–1226
24. Yu DRC (2004) Watershed segmentation for breast tumor in 2D sonography. Ultrasound Med Biol 30:625–632
25. Zhao N, Tong N, Ruan D, Sheng K (2019) Fully automated pancreas segmentation with two-stage 3D convolutional neural networks. In: International conference on medical image computing and computer assisted intervention (MICCAI)
26. Zhou Z, Wu W, Wu S, Tsui PH, Lin CC, Zhang L, Wang T (2014) Semi-automatic breast ultrasound image segmentation based on mean shift and graph cuts. Ultrason Imaging 36(4):256–276