Automated Segmentation of Lesions in Ultrasound Using Semi-pixel-wise Cycle Generative Adversarial Nets

Jie Xing, Zheren Li, Biyuan Wang, Bingbin Yu, Farhad G. Zanjani, Aiwen Zheng, Remco Duits, Tao Tan

arXiv:1905.01902v1 [cs.CV] 6 May 2019

***Abstract*— Breast cancer is the most common invasive cancer with the highest cancer occurrence in females. Handheld ultrasound is one of the most efficient ways to identify and diagnose the breast cancer. The area and the shape information of a lesion is very helpful for clinicians to make diagnostic decisions. In this study we propose a new deep-learning scheme, semi-pixel-wise cycle generative adversarial net (SPCGAN) for segmenting the lesion in 2D ultrasound. The method takes the advantage of a fully connected convolutional neural network (FCN) and a generative adversarial net to segment a lesion by using prior knowledge. We compared the proposed method to a fully connected neural network and the level set segmentation method on a test dataset consisting of 32 malignant lesions and 109 benign lesions. Our proposed method achieved a Dice similarity coefficient (DSC) of 0.92 while FCN and the level set achieved 0.90 and 0.79 respectively. Particularly, for malignant lesions, our method increases the DSC (0.90) of the fully connected neural network to 0.93 significantly ($p < 0.001$). The results show that our SPCGAN can obtain robust segmentation results and may be used to relieve the radiologists' burden for annotation.**

***Index Terms*— Lesion Segmentation, Deep Learning, Generative Adversarial Networks, Breast Cancer, Ultrasound Image Analysis**

I. Introduction

Breast cancer is one of the leading causes of death for women in the UK. According to the statistics published by Cancer Research UK, there are about 155 women in 100,000 suffering from breast cancer in the UK and incidence rate is around 10% for females in other European countries while this number is over 12.5% for breast cancer with the American females [5]. Since the causes of breast cancer still remain unknown, early diagnosis of breast cancer plays a significant role in reducing the death rate and maintaining the quality of the life after treatments [13].

Ultrasound imaging technology has developed rapidly in recent years. Compared to mammography, there is no radiation damage to women from ultrasound imaging. It is easy to obtain any cross-sectional images of breast tissue by manipulating the handheld ultrasound while normally only two projections are obtained from mammography. It provides an easy way to assess if a lesion is solid or fluid-filled [18]. Ultrasound detects early-stage cancers in women with mammography-negative dense breasts, with higher contribution in women younger than 50 years [3]. Moreover, breast ultrasound is simple, effective and low cost. Because of all these advantages, it can be applied in a large scale for imaging, for example, in China [28].

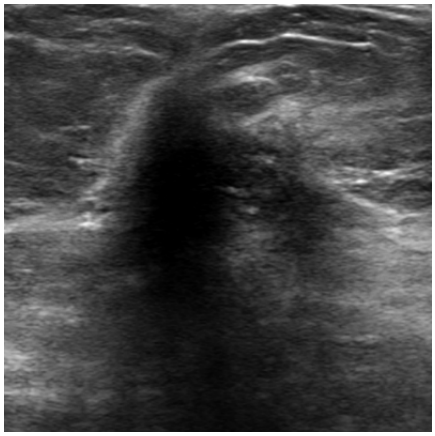


Fig. 1. A malignant lesion in breast ultrasound

In the clinical workflow of breast ultrasound imaging, radiologists often report the sizes of breast lesions, describe the lesions according to BI-RADS lexicon [6] and estimate the final BI-RADS score. An accurate delineation of breast lesion can help radiologists describe margin, shape and posterior features. However, manual segmentation of breast lesions is time-consuming and tedious. The segmentation also varies from one reader to another. Therefore, the automated segmentation can play a key role in facilitating reporting. In terms of detection and diagnosis, computers can also assist radiologists to make decisions that improve the effectiveness of ultrasound reading. For example, computer techniques [15], [19], [32]–[35] have been proposed to delineate the contour of lesions or directly detect or diagnose breast lesions. Most of these computer-aided diagnoses or detections include a module of segmentation. Therefore, it is important to develop a robust and accurate segmentation method.

Breast lesion segmentation is very challenging, especially when there is the presence of noise, the ill-defined edges, irregular shapes, and different posterior behaviors of lesions. As Fig. 1 shows, there is strong shadowing in the posterior and upper region, the lesion boundary is fuzzy and not clear. Therefore, there is a risk that segmentation algorithms fail, causing oversegmentation.

There are two types of segmentation methods: contour-based and region-based method. The contour-based segmentation relies on finding the optimal contour to enclose the whole breast lesion. Region-based method aims at assigning label to every image pixel. Jing et al. [4] proposed an iterative segmentation scheme to refine the initial contour and perform self-examination and correction on the segmentation result. Their best intersection of the computer and the reference segmented area was 0.84. Tan et al. [31] proposed a novel depth-dependent dynamic programming technique and obtained a Dice similarity coefficient (DSC) of 0.73. This accuracy was then improved by Kozegar et al. with a specific level set algorithm [14]. Horsch et al. [9] presented a computationally efficient segmentation

Jie Xing, Zheren Li have equal contribution
J. Xing and Ai-Wen Zheng are with Zhejiang cancer hospital China
Z. Li is with Department of Bioengineering, Imperial College London, London SW7 2AZ,UK
T. Tan and R. Duits are with Department of Mathematics and Computer Science, Eindhoven University of Technology, Eindhoven 5600 MB, The Netherlands
Farhad G. Zanjani is with Department of Electrical Engineering, Eindhoven University of Technology, Eindhoven 5600 MB, The Netherlands
T. Tan (t.tan1@tue.nl) is the corresponding author
B. Wang is with Department of Computing, Tokyo Institute of Technology, Tokyo, Japan
B.Yu is with Robotic Innovation Center, German Research Center of Artificial Intelligence, Bremen, Germany

algorithm for breast masses on sonography, which is based on maximizing a utility function over partition margins defined through gray value thresholding of a preprocessed image. Their algorithm was evaluated on a database of 400 cases and the reported average overlap rate was 0.73. The challenge of applying contour-based method is to make sure the contour evolvement is not trapped by non-breast edges. For region-based segmentation, both traditional methods and machine-learning-based pixel classification methods were investigated. Feng et al. adopted an adaptive fuzzy C-means algorithm and the obtained DSC is 0.925. Pons et al. [24]reported that their evaluated automated method achieved a DSC of 0.49 using a Markov Random Field (MRF) and a Maximum a Posteriori (MAP) approach, by applying it to clinical data. Agarwal et al. [1] developed a semi-automatic framework for breast lesion segmentation in ABUS volumes which is based on the Watershed algorithm. Rodrigues et al. [26] took the advantage of pixel-wise classification and achieved a DSC of 0.824. Kumar et al. [17] proposed convolutional neural network approaches for breast ultrasound lesion segmentation and the computer tool effectively segmented the breast masses, achieving a mean DSC of 0.82.

As one branch of machine learning, deep learning has become popular as a self-taught approach in which features are computed in an automatic manner instead of combining manually designed features [8], [10]–[12], [16], [30]. These approaches have rapidly become state-of-the-art that outperform other traditional methods in the segmentation tasks with ultrasound. There are generally two ways of applying deep learning: patch-wise classification using convolutional neural networks (CNN) and pixel-wise classification using fully convolutional networks (FCN) such as U-Net architecture [27]. These techniques have gained propitiatory for the segmentation tasks. Most existing deep learning based methods still rely on image information (lesion boundaries) while the prior knowledge of breast lesion shape is not well used, although they have already obtained accurate segmentation results. To further improve classification results, proper incorporation of prior knowledge is necessary. In this work we use a model which tend to learn prior knowledge of breast lesions and is able to properly deal with fuzziness or even the absence of a visible lesion edge in the some sections of lesions. We proposed a generative adversarial net (GAN) based framework, semi-pixel-wise cycle generative adversarial net (SPCGAN), for segmenting the lesion in 2D ultrasound. The main contributions of this paper can be concluded as follows: This is the first research for breast lesion segmentation task by combing the GAN loss and FCN loss. The combination will make the segmentation not only reply on lesion boundary but also mimic the way of human annotation. Because of the power of GAN, the scheme requires less data to train a model with effectiveness and robustness.

II. Methods

A. Semi-pixel-wise Cycle Generative Adversarial Net

Generative adversarial net (GAN) is a framework which consists of two models for the estimation of generative results via an adversarial process [7]. There is a generative model G which tries to produce data that is similarly distributed as training data. Simultaneously, a discriminative model D evaluates the authenticity of a sample data coming from training set. Both G and D aim to deteriorate the performance of each other, therefore the loss function of GANs mimics a two-player minimax game and is expressed as follows:

$$\min_G \max_D \mathcal{L}(D,G) = \mathbb{E}_{\boldsymbol{x}\sim p_{data}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1-D(G(\boldsymbol{z}))], \tag{1}$$

where $p_{\boldsymbol{z}}(\boldsymbol{z})$ is the distribution of training data and $G(\boldsymbol{z})$ is the generative distribution over training data $\boldsymbol{x}$. $D(\boldsymbol{x})$ gives the probability that data $\boldsymbol{x}$ belongs to the training data rather than $G(\boldsymbol{z})$. D is trained to maximize the probability to assign correct labels to the input data. Meanwhile, G is trained to disturb D's judgment, which is to minimize $\log(1-D(G(\boldsymbol{z}))$.

In a CycleGAN model [40], the generator no longer generates data from random source such as white noise. There are two target domains which can be unpaired for data transfer between each other and the data generation process is now drawn an analogy to an autoencoder. There are two generators to translate data in one domain to the other, which can be regarded as an encoder and decoder respectively. There are also two discriminators and each of them tries to discriminate the authenticity of the data that belongs to the corresponding domain. The complete generation and discrimination process forms a cycle as shown in Figure 2.

Comparing to transfer data between target domains via two GANs, the cycle mechanism of CycleGAN network necessarily guarantees the one-one mapping relationship between the input and output data and therefore rules out the possibility that any input data can be mapped to induce a set of output data distributions which match the target domain [40]. The adversarial loss function of a CycleGAN is modified as follows:

$$\mathcal{L}_{GAN}(G,D_Y,X,Y) = \mathbb{E}_{\boldsymbol{y}\sim p_{data}(\boldsymbol{y})}[\log D_Y(\boldsymbol{y})] + \mathbb{E}_{\boldsymbol{x}\sim p_{\boldsymbol{data}}(\boldsymbol{x})}[\log(1-D_Y(G(\boldsymbol{x}))], \tag{2}$$

where $G(x)$ is the generator takes the input from domain X and tries to generate data of similar distribution as in domain Y. D_Y aims to distinguish between generated samples $G(x)$ and real samples y from domain Y while G tries to minimize this objective. To ensure the expected mapping from input to the desired output, there is also a cycle loss to evaluate the decoder performance, which is to check whether the translated data can be brought back to the original domain, i.e., $x \to G_{xy}(x) \to G_{yx}(G_{xy}(x)) \approx x$ and vice versa in the translation from domain Y to domain X. This forms a cycle consistency in both forward and backward directions. The loss function is expressed as follows:

$$\mathcal{L}_{cyc}(G_{xy},G_{yx}) = \mathbb{E}_{\boldsymbol{x}\sim p_{data}(\boldsymbol{x})}[||G_{yx}(G_{xy}(x))-x||_1] + \mathbb{E}_{\boldsymbol{y}\sim p_{data}(\boldsymbol{y})}[||G_{xy}(G_{yx}(y))-y||_1]. \tag{3}$$

In this research, we adopted the general architecture of CycleGAN to our model but manipulated on the discriminator by adding an extra loss related to the pixel-wise classification. Our generator is an FCN which only contains convolutional layers. FCN can be trained end-to-end by upsampling and deconvoluting the feature maps extracted from convolutional layers and output pixel-wise classifications on the input images. It is widely used in semantic segmentations [20].

The discriminator of our model is modified to evaluate the adversarial loss between the manually annotated segmentation ground truth and the generated segmentation in every pixel, rather than the probability of classification on the whole image. This pixel-wise loss is calculated by the following formula:

$$\mathcal{L}_{\text{pixel-wiseGAN}}(G,D) = \sum_{i=1}^{N} \frac{\|\boldsymbol{G_i}-\boldsymbol{D_i}\|_2^2}{N}, \tag{4}$$

where G is the generated segmentation image and D is the ground truth segmentation. For each pixel pair between the two segmentations, we calculate the difference to obtain the pixel-wise loss and sum up over the whole image then divided by the total number of pixels N to obtain the average. G_i and D_i are the i-th pixel in the generated segmentation and ground truth respectively. In the forward

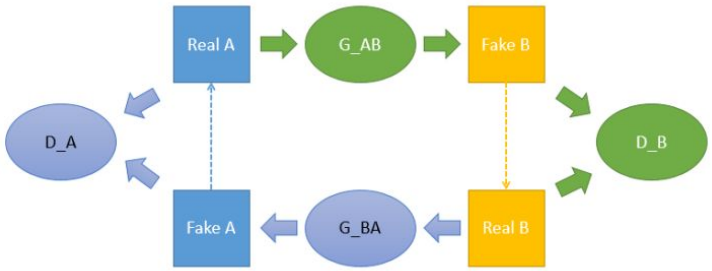


Fig. 2. The architecture of CycleGAN model: forward GAN on the upper half is aiming to translate image from Domain A to Domain B via generator G_AB and being distinguished by discriminator D_B from Domain B. Backward GAN on the lower half is working in the opposite way to take in image in Domain B and translate into Domain A. D_A distinguishes the generated image in Domain A from the real one.

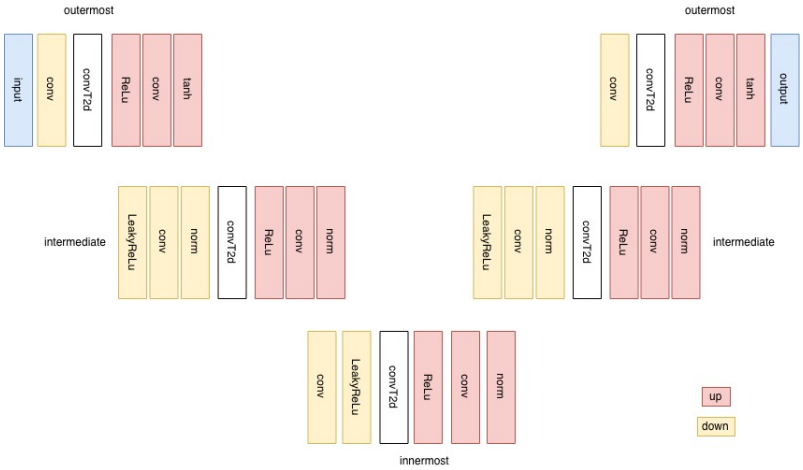


Fig. 3. The architecture of the generator in our model. It is an FCN based model and has the structure of U-Net. The yellow blocks represent the upper layers and the pink blocks represent the lower layers in the U shape structure.

process of the cycle, the discriminator tries to defy and reject the generated segmentation. While the generator would like to minimize the pixel-wise loss so that makes the generated segmentations get accepted by the discriminator. The architecture of our model is shown in Figure4:

In the algorithm of GAN, the input data z is drawn from a simple prior probability distribution such as Gaussian. Therefore, z is essentially a latent vector of unstructured noise. The probability distribution of z in the latent space is then aligned to the real data distribution. It is not able to take the advantage of the prior knowledge of images. However, in a CycleGAN-like algorithm, because there are two domains, we could draw the prior knowledge from the source image as the prior distribution and then generate samples based on it rather than imposing a random probability. In our implementation, the prior knowledge of annotated segmentation of breast lesion is learned by the model so that it is able to properly deal with images with ambiguous features, for example, in the absence of visible lesion edge.

B. Fully Convolutional Network (FCN)

The generator used above itself is an FCN and this type of network is applied extensively in ultrasound [39] [38] [21], [23], [25], [37] [29] [36] for various applications. To show the advantage of our model, we also applied FCN (Fig. 3) for comparison. The network is exactly the same as the generator of our proposed model.

C. The Level Set Method

To compare our deep learning scheme with a traditional segmentation method, we applied a geodesic-active-contour (GAC) based level set method. This level set method is with curvature and advection terms introduced by Caselles et al. [2]. The partial differential equation describing the motion of the contour is defined by:

$$\Phi_t + g \cdot (1 + \varepsilon k)|\nabla\Phi| + \alpha \nabla g \cdot \nabla\Phi = 0 \tag{5}$$

with Φ the level set function, ε the curvature influence, k the curvature, α the advection influence and g the gradient based speed function:

$$g(I) = 1/(1 + |(\nabla * G)(I)|), \tag{6}$$

where I is the image intensity and $(\nabla * G)$ is the derivative of Gaussian operator.

We initialize the level set with the center of the lesion. The segmentation can be controlled by setting the weights (α, ε) of propagation, curvature and advection term. The propagation term controls the inflation or balloon force of the segmentation, the curvature controls the smoothness of the segmentation and the advection term in the update equation behaves like a doublet and attracts the contour to the lesion edge [31].

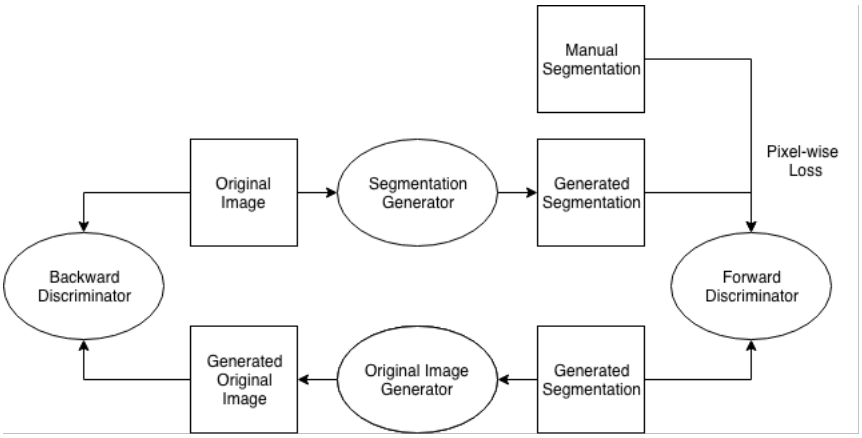


Fig. 4. The architecture of SPCGAN with pixel-wise loss. The pixel-wise loss is only applied in the forward cycle.

III. Materials

A. Datasets

This study is based on 2D BUS DICOM images of abnormal patients. All DICOM images were scanned from SIEMENS MED SMS USG S2000 and TOSHIBA Aplio400 TUS-A400 Ultrasound System. For this study, we collected a dataset of 670 breast lesion ultrasound images from different women (aged 18-70) that had no history of breast cancer. Among the 670 images, 640 were scanned from SIEMENS Ultrasound System and 30 were scanned from TOSHIBA Ultrasound System. If there were a number of DICOM images from the same lesion, the DICOM image containing the maximum area of lesion was collected into the dataset. The type of lesion has been clinically diagnosed as malignant or benign. Among the 640 lesions from SIEMENS Ultrasound System, 120 are malignant lesions and 520 are benign lesions.

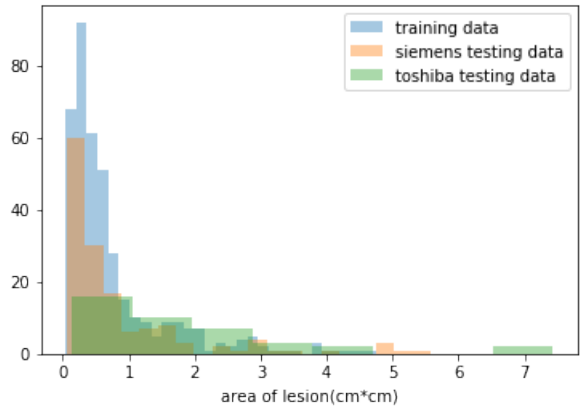


Fig. 5. The histogram of lesion areas in training and testing dataset.

In our study dataset, for SIEMENS images, 399 were used in the training phase, 141 were used in the testing phase and 100 were used in the validation phase. All TOSHIBA images were used for testing the generality of the model trained by SIEMENS images. During the image preprocessing stage, the original DICOM images were re-sampled to 0.1mm spacing in both horizontal and vertical directions. After that the ROI images were cropped from the re-sampled DICOM images with a size of 400*400 for easy processing, and the center of the lesions were the center of the ROI images.

For each ROI image, we manually generated reference lesion segmentations by using a MATLAB program. These manual segmentations were performed by an experienced researcher with 10 years of experience in breast ultrasound.

B. Performance Evaluations

In this study, Dice similarity coefficient is used for describing the accuracy of the segmentation by different methods. Dice similarity coefficient is a statistic used for comparing the similarity between two samples, and is defined as follows:

$$DSC = \frac{2\,|X \cap Y|}{|X| + |Y|}, \tag{7}$$

where $|X|$ is the area of lesion by manual segmentation and $|Y|$ is the area of lesion segmented by automatic methods. The larger the Dice similarity coefficient is, the higher accuracy of the computational segmentation is. It ranges between 0 and 1.

C. Implementation platform

Our network was trained on a workstation equipped with an NVIDIA GeForce GTX 1080Ti GPU.

IV. Experiments and Results

A. Statistical Analysis

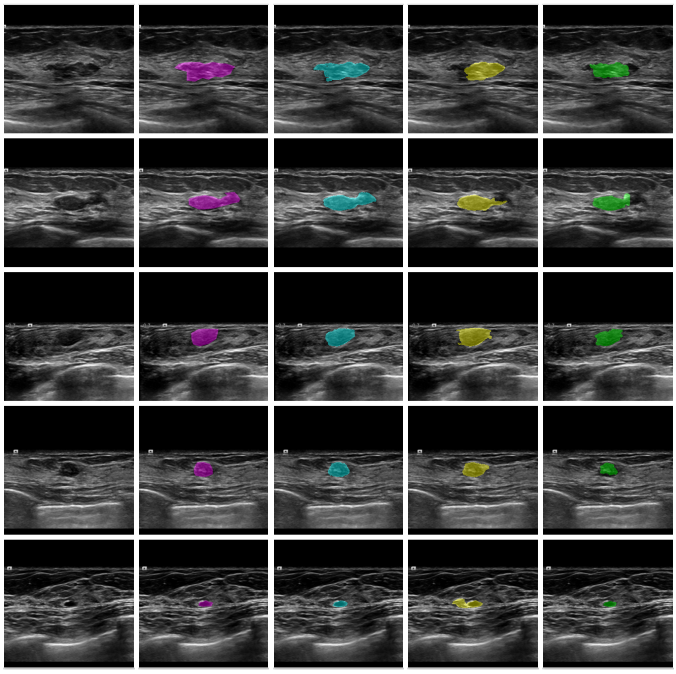
One-sided paired t-tests are used for statistical analysis when comparing results from different segmentation methods. The hypothesis in this study will be tested to control type I error rate at alpha = 0.05. The hypothesis is that the DSC of the SPCGAN is superior to (bigger than) that of U-Net or the level set method, for statistical significance level alpha = 0.05.

B. Comparisons among SPCGAN, U-Net and Level Set

TABLE I
DSC of different segmentation methods.

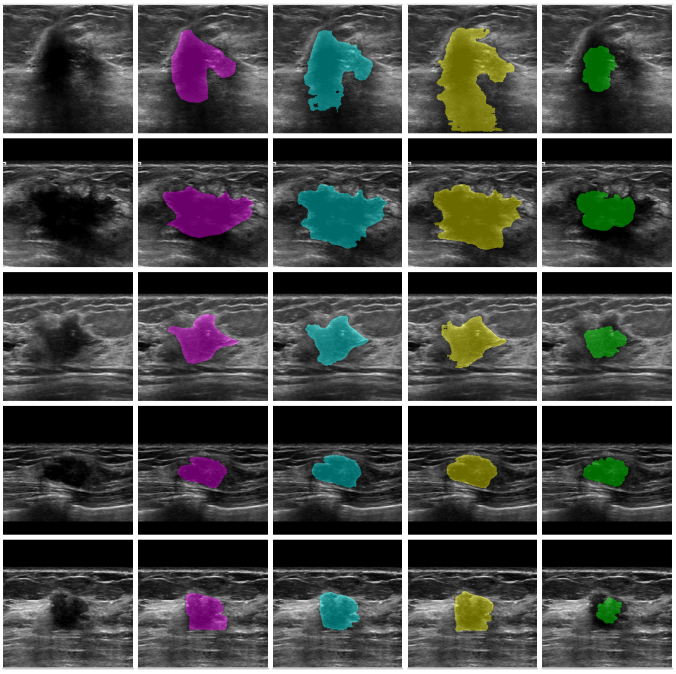
	SPCGAN	U-Net	level set
All lesions	0.92±0.04	0.90±0.07	0.79±0.17
benign	0.92±0.04	0.90±0.07	0.83±0.16
malignant	0.93±0.04	0.90±0.07	0.65±0.17

To evaluate the effect of SPCGAN framework on the breast ultrasound lesion segmentation accuracy, we compared it with U-Net framework and the traditional segmentation method level set. Table I summarizes the DSC of different methods on our test database of



(a) (b) (c) (d) (e)

Fig. 6. Comparison of SPCGAN and other segmentation methods of benign lesions. (a) shows original image of benign lesions, (b) shows the manual annotation, (c) shows the result of SPCGAN ,(d) and (e) show results from U-Net and level set.



(a) (b) (c) (d) (e)

Fig. 7. Comparison of SPCGAN and other segmentation methods of malignant lesions.(a) shows original image of malignant lesions, (b) shows the manual annotation, (c) shows the result of SPCGAN ,(d) and (f) show results from U-Net and the level set method.

141 lesions from the entire dataset and 32 lesions are malignant. DSC values were obtained from 109 benign and 32 malignant breast lesions. Comparing the overall results of 3 different methods, we see that SPCGAN performed better by 2% improvement compared to the U-Net method ($p<0.001$). The traditional method level set performed significantly worse compared to SPCGAN method ($p<0.001$). Furthermore, compared to the traditional segmentation method level set, the DSCs obtained from SPCGAN and U-Net still remain high no matter a lesion is benign or malignant. The DSCs of malignant lesions from the level set method were much lower than the DSCs of benign lesions.

Fig.6 displays the segmentation results of our SPCGAN, U-Net and the level set method from benign lesions. Compared with the U-Net (d) and the level set (e) method, the results of our SPCGAN (c) show good agreements with the manual contours of the lesions. The segmentations from SPCGAN are very close to manual segmentations.

The examples given in Fig.7 correspond to the segmentation results of our SPCGAN, U-Net and the level set method from malignant lesions. The U-Net tends to oversegment the cancer when there is posterior shadowing, especially for the lesion in the first row. SPCGAN shows relatively more robust performance compared to U-Net and the level set method.

Fig.8 illustrates boxplots of DSC for different segmentation methods. We can see that results from our SPCGAN have much less variance compared to other methods.

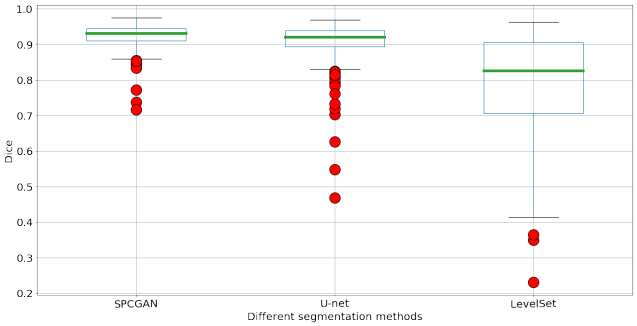


Fig. 8. Boxplot of DSC from different segmentation methods.

C. Comparisons between Models Trained with Varying Number of Samples

In order to compare the performance of SPCGAN and U-Net model, we use varying numbers of samples to train the model and then test with the same testing dataset. The changes in the performance of SPCGAN and U-Net when trained with 20, 40, 80, 200 and 399 samples are displayed in Fig. 9.

From Fig. 9 and Table II, we can observe that SPCGAN model obtained better results among all training sample numbers, especially when sample size was small. Although U-Net model trained with 200 samples can achieve an average DSC of 0.90, it is still 0.01 lower than SPCGAN. Particularly, when training samples increased to 399, the DSC of SPCGAN improved to 92% while that of U-Net remained with 90%.

Fig. 10 displays one case for which segmentation was performed by SPCGAN and U-Net trained with varying numbers of samples. This example demonstrates how unclear boundary and shadow in ultrasound images may affect segmentation algorithms.

D. Results from Test Data from Other Manufactures

To explore the performance of SPCGAN on the segmentation quality with test data from different manufactures, we collected 30 BUS images of breast disease scanned from TOSHIBA Ultrasound

TABLE II
DSC OF DIFFERENT SEGMENTATION METHODS.

number of training samples	SPCGAN	U-Net
20	0.79±0.26	0.64±0.37
40	0.85±0.20	0.83±0.23
80	0.88±0.10	0.87±0.15
200	0.91±0.08	0.90±0.09
399	0.92±0.04	0.90±0.07

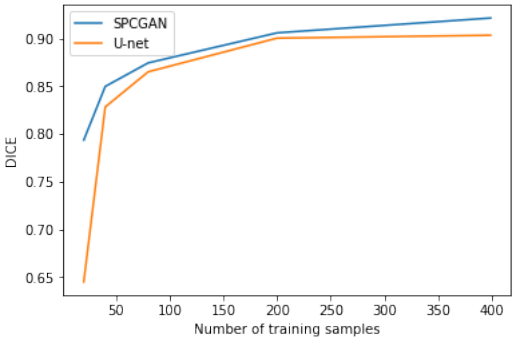


Fig. 9. DSC values obtained when SPCGAN and U-Net were trained with different number of training samples.

System and applied our model which was trained on SIEMENS images only.

From Table III, we can observe that the difference between DSC values of SPCGAN and U-Net was not statistically significant (p=0.14 with paired t-test). The example given in Fig. 11 corresponds to a breast cancer with the ill-defined boundary. They are both very robust but the mean DSC from SPCGAN is still higher.

V. CONCLUSION AND DISCUSSION

In this study, we proposed a CycleGAN based model for segmenting breast lesions in 2D breast ultrasound. We compared our model with FCN and the level set based approach. The results show that our model is the most robust and accurate with a DSC of 0.92± 0.04.

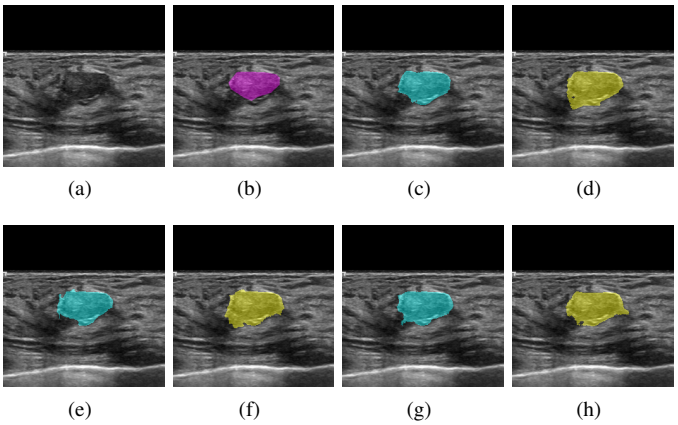


Fig. 10. A breast lesion, which was segmented by SPCGAN and U-Net trained with varying number of samples. (a) shows the original image, (b) shows the manual annotation, (c) shows the result of SPCGAN with 80 training samples,(d) shows the result of U-Net with 80 training samples, (e) shows the result of SPCGAN with 200 training samples, (e) shows the result of U-Net with 200 training samples, (g) and (h) show result of SPCGAN and U-Net with 399 training samples respectively.

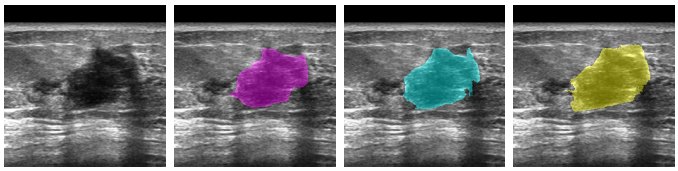


Fig. 11. Comparison between SPCGAN and U-Net in a TOSHIBA image. (a) shows original image of benign lesions, (b) shows the manual annotation, (c) shows the result of SPCGAN ,and (d) shows results from U-Net.

TABLE III
DICE VALUES FOR TOSHIBA IMAGES

	SPCGAN	U-Net
30 test images	0.93±0.02	0.92±0.04

Without retraining, the same model is applied on ultrasound images from a different manufacture, resulting a DSC of 0.93± 0.02.

The novelty of our work is the combination of the CycleGAN and the pixel wise cost which makes the model has the advantage of both GAN and FCN. However, for some challenging images, especially when calcification is present, the segmentation is still not very smooth. From Figure 9, we can observe an improvement on DSC from 0.79 to 0.92 when the number of training images is increased. In the future, we will increase the size of our dataset and eventually our model will be more robust. Another approach is to apply post processing, for example, Markov random filtering, to make segmentation smooth and complete. Both deep learning based methods are significantly better than the traditional level set approach in both malignant and benign cases, even when the deep learning model is only trained with 20% percent of the dataset. With more annotated data, the performance of supervised learning can be improved significantly. It is still possible to enhance the segmentation performance by combining deep learning and the level set method together. Researchers [22] showed that with a combined approach, they achieved the most accurate results in the semi-automated problem. In the future, we will investigate more possibilities.

In the cyclic training process of our model, the forward generator firstly produces automatic segmentation of the breast lesions indistinguishable to the forward discriminator by minimizing the adversarial loss, cycle consistency loss and pixel-wise cost. This generated segmentation is then fed into the backward generator which tries to get it recover to the original image. During this stage, the pixel-wise cost is no longer applied as it is hard to recover the original image from the segmentation in regards of pixel level. The implementation of CycleGAN algorithm effectively utilizes the prior knowledge of lesion images to provide the prior distribution in the latent space for the input. By imposing this prior probability distribution, the mapping between input data sample and real data is under a more sensible constraint. In ultrasound annotation tasks, the segmentation is challenging because of poor quality, posterior shadowing and weak boundaries. In this case, the use of prior knowledge would make segmentation robust.

Accurate segmentation would help describe shape, orientation, margins, echo pattern, posterior acoustic features, and surrounding tissue alterations of a lesion in BI-RADS US lexicon. The description would also aid radiologists or computer algorithms [33], [34] to diagnose a lesion. Given accurate segmentations, it would be logical to design further deep learning networks to differentiate malignant lesions from benign lesions or generate BI-RADS scores in ultra-

sound.

One limitation of our study is that we compared the segmentations results to annotations from only one reader. It is also interesting to show the comparison of DSCs between two readers as a reference. Moreover, as the CycleGAN model has the ability of learning prior knowledge, for example the shape, it is possible that it only learns the style of one reader. Whether the prior knowledge from one reader is sufficient to obtain good inter-reader variability shall be investigated in the feature.

In the real clinical practice, there are ultrasound devices from different manufacturers deployed. These images varies in resolution, contrast, and the presence of noise. In this study, we evaluated the possibility of applying our deep learning model trained on SIEMENS ultrasound images only to TOSHIBA ultrasound images. The segmentation accuracy still remains high. Researchers can focus on developing robust algorithms on different types of ultrasound images, which is important to make computer techniques available to real world practice.

In summary, we proposed a robust framework for the purpose of lesion segmentation in breast ultrasound and the segmentation accuracy is comparable to human annotations. Our method has the potential to help radiologists delineate breast lesion and improve the efficiency of workflow for reporting and inter-/intra- reader variability. While focused on one type of lesions in ultrasound, future work can address a wider range of objects in different medical images.

References

[1] Richa Agarwal, Oliver Diaz, Xavier Llad??, Albert Gubern-M??rida, Joan C Vilanova, and Robert Mart?? Lesion segmentation in automated 3d breast ultrasound: Volumetric analysis. *Ultrasonic imaging*, 40:97–112, March 2018.

[2] Vicent Caselles, Ron Kimmel, and Guillermo Sapiro. Geodesic active contours. *International Journal of Computer Vision*, 22(1):61, February 1997.

[3] Vittorio Corsetti, Nehmat Houssami, Aurora Ferrari, Marco Ghirardi, Sergio Bellarosa, Osvaldo Angelini, Claudio Bani, Pasquale Sardo, Giuseppe Remida, Enzo Galligioni, et al. Breast screening with ultrasound in women with mammography-negative dense breasts: evidence on incremental cancer detection and false positives, and associated cost. *European journal of cancer*, 44(4):539–544, 2008.

[4] Jing Cui, Berkman Sahiner, Heang-Ping Chan, Alexis Nees, Chintana Paramagul, Lubomir M Hadjiiski, Chuan Zhou, and Jiazheng Shi. A new automated method for the segmentation and characterization of breast masses on ultrasound images. *Medical physics*, 36:1553–1565, May 2009.

[5] C DeSantis, R Siegel, and A Jemal. Breast cancer facts & figures 2015-2016. atlanta: American cancer society. *Inc*, 2015:1–44, 2015.

[6] Carl J D'Orsi. *ACR BI-RADS atlas: breast imaging reporting and data system*. American College of Radiology, 2013.

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European Conference on Computer Vision*, pages 630–645, 2016.

[9] K Horsch, M L Giger, L A Venta, and C J Vyborny. Automatic segmentation of breast lesions on ultrasound. *Medical physics*, 28:1652–1659, August 2001.

[10] Gao Huang, Zhuang Liu, Kilian Q Weinberger, and Laurens van der Maaten. Densely connected convolutional networks. *arXiv preprint arXiv:1608.06993*, 2016.

[11] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456, 2015.

[12] Dongsheng Jiang, Weiqiang Dou, Luc Vosters, Xiayu Xu, Yue Sun, and Tao Tan. Denoising of 3d magnetic resonance images with multi-channel residual learning of convolutional neural network. *Japanese journal of radiology*, 36(9):566–574, 2018.

[13] Kamlesh Kaul and Fabienne Marie-Louise Daguilh. Early detection of breast cancer: is mammography enough? *Hospital Physician*, 38(9):49–54, 2002.

[14] E Kozegar, M Soryani, H Behnam, M Salamati, and T Tan. Mass segmentation in automated 3-d breast ultrasound using adaptive region growing and supervised edge-based deformable model. *IEEE transactions on medical imaging*, 37:918–928, April 2018.

[15] Ehsan Kozegar, Mohsen Soryani, Hamid Behnam, Masoumeh Salamati, and Tao Tan. Breast cancer detection in automated 3d breast ultrasound using iso-contours and cascaded rusboosts. *Ultrasonics*, 79:68–80, 2017.

[16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Commun. ACM*, 60(6):84–90, May 2017.

[17] Viksit Kumar, Jeremy M Webb, Adriana Gregory, Max Denis, Duane D Meixner, Mahdi Bayat, Dana H Whaley, Mostafa Fatemi, and Azra Alizad. Automated and real-time segmentation of suspicious breast masses using convolutional neural network. *PloS one*, 13:e0195816, 2018.

[18] H. Laine, J. Rainio, H. Arko, and T. Tukeva. Comparison of breast structure and findings by X-ray mammography, ultrasound, cytology and histology: A retrospective study. *European Journal of Ultrasound*, 2(2):107–115, April 1995.

[19] Haixia Liu, Tao Tan, Jan van Zelst, Ritse Mann, Nico Karssemeijer, and Bram Platel. Incorporating texture features in a computer-aided breast lesion diagnosis system for automated three-dimensional breast ultrasound. *Journal of Medical Imaging*, 1(2):024501, 2014.

[20] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[21] Fausto Milletari, Seyed-Ahmad Ahmadi, Christine Kroll, Annika Plate, Verena Rozanski, Juliana Maiostre, Johannes Levin, Olaf Dietrich, Birgit Ertl-Wagner, Kai B?tzel, and Nassir Navab. Hough-CNN: Deep learning for segmentation of deep brain regions in MRI and ultrasound. *Computer Vision and Image Understanding*, 164:92–102, November 2017.

[22] Tuan Anh Ngo, Zhi Lu, and Gustavo Carneiro. Combining deep learning and level set for the automated segmentation of the left ventricle of the heart from cardiac cine magnetic resonance. *Medical image analysis*, 35:159–171, 2017.

[23] Ozan Oktay, Enzo Ferrante, Konstantinos Kamnitsas, Mattias Heinrich, Wenjia Bai, Jose Caballero, Stuart A Cook, Antonio de Marvao, Timothy Dawes, Declan P O???Regan, et al. Anatomically constrained neural networks (acnns): application to cardiac image enhancement and segmentation. *IEEE transactions on medical imaging*, 37(2):384–395, 2018.

[24] Gerard Pons, Joan Mart??, Robert Mart??, Sergi Ganau, and J Alison Noble. Breast-lesion segmentation combining b-mode and elastography ultrasound. *Ultrasonic imaging*, 38:209–224, May 2016.

[25] Hariharan Ravishankar, S Thiruvenkadam, R Venkataramani, and V Vaidya. Joint deep learning of foreground, background and shape for robust contextual segmentation. In *International Conference on Information Processing in Medical Imaging*, pages 622–632. Springer, 2017.

[26] Rafael Rodrigues, Rui Braz, Manuela Pereira, Jos?? Moutinho, and Antonio M G Pinheiro. A two-step segmentation method for breast ultrasound masses based on multi-resolution analysis. *Ultrasound in medicine & biology*, 41:1737–1748, June 2015.

[27] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[28] Qing-Kun Song, Xiao-Li Wang, Xin-Na Zhou, Hua-Bing Yang, Yu-Chen Li, Jiang-Ping Wu, Jun Ren, and Herbert Kim Lyerly. Breast cancer challenges and screening in china: lessons from current registry data and population screening studies. *The oncologist*, 20(7):773–779, 2015.

[29] V. Sundaresan, C. P. Bridge, C. Ioannou, and J. A. Noble. Automated characterization of the fetal heart in ultrasound images using fully convolutional neural networks. In *Proc. IEEE 14th Int. Symp. Biomedical Imaging (ISBI 2017)*, pages 671–674, April 2017.

[30] T. Tan, Z. Li, H. Liu, F. G. Zanjani, Q. Ouyang, Y. Tang, Z. Hu, and Q. Li. Optimize transfer learning for lung diseases in bronchoscopy using a new concept: Sequential fine-tuning. *IEEE Journal of Translational Engineering in Health and Medicine*, 6:1–8, 2018.

[31] Tao Tan, Albert Gubern-M??rida, Cristina Borelli, Rashindra Manniesing, Jan van Zelst, Lei Wang, Wei Zhang, Bram Platel, Ritse M Mann, and Nico Karssemeijer. Segmentation of malignant lesions in 3d breast ultrasound using a depth-dependent model. *Medical physics*, 43:4074, July 2016.

[32] Tao Tan, Jan-Jurre Mordang, Jan Zelst, André Grivegnée, Albert Gubern-Mérida, Jaime Melendez, Ritse M Mann, Wei Zhang, Bram Platel, and Nico Karssemeijer. Computer-aided detection of breast cancers using haar-like features in automated 3d breast ultrasound. *Medical physics*, 42(4):1498–1504, 2015.

[33] Tao Tan, Bram Platel, Henkjan Huisman, Clara I Sánchez, Roel Mus, and Nico Karssemeijer. Computer-aided lesion diagnosis in automated 3-d breast ultrasound using coronal spiculation. *IEEE transactions on medical imaging*, 31(5):1034–1042, 2012.

[34] Tao Tan, Bram Platel, Roel Mus, Laszlo Tabar, Ritse M Mann, and Nico Karssemeijer. Computer-aided detection of cancer in automated 3-d breast ultrasound. *IEEE transactions on medical imaging*, 32(9):1698–1706, 2013.

[35] Tao Tan, Bram Platel, Thorsten Twellmann, Guido van Schie, Roel Mus, André Grivegnée, Ritse M Mann, and Nico Karssemeijer. Evaluation of the effect of computer-aided classification of benign and malignant lesions on reader performance in automated three-dimensional breast ultrasound. *Academic radiology*, 20(11):1381–1388, 2013.

[36] L. Wu, Y. Xin, S. Li, T. Wang, P. A. Heng, and D. Ni. Cascaded fully convolutional networks for automatic prenatal ultrasound image segmentation. In *Proc. IEEE 14th Int. Symp. Biomedical Imaging (ISBI 2017)*, pages 663–666, April 2017.

[37] Lingyun Wu, Yang Xin, Shengli Li, Tianfu Wang, Pheng-Ann Heng, and Dong Ni. Cascaded fully convolutional networks for automatic prenatal ultrasound image segmentation. In *Biomedical Imaging (ISBI 2017), 2017 IEEE 14th International Symposium on*, pages 663–666. IEEE, 2017.

[38] Moi Hoon Yap, Gerard Pons, Joan Martí, Sergi Ganau, Melcior Sentís, Reyer Zwiggelaar, Adrian K Davison, and Robert Martí. Automated breast ultrasound lesions detection using convolutional neural networks. *IEEE journal of biomedical and health informatics*, 2017.

[39] Yizhe Zhang, Michael TC Ying, Lin Yang, Anil T Ahuja, and Danny Z Chen. Coarse-to-fine stacked fully convolutional nets for lymph node segmentation in ultrasound images. In *Bioinformatics and Biomedicine (BIBM), 2016 IEEE International Conference on*, pages 443–448. IEEE, 2016.

[40] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.