Attention Enriched Deep Learning Model for Breast Tumor Segmentation in Ultrasound Images

Aleksandar Vakanski\textsuperscript{a}, Min Xian\textsuperscript{a}, and Phoebe E. Freer\textsuperscript{b}

\textsuperscript{a} Department of Computer Science, University of Idaho, Idaho Falls, USA
\textsuperscript{b} University of Utah School of Medicine, Salt Lake City, USA

\textsuperscript{a}Corresponding Author: Aleksandar Vakanski, 1776 Science Center Drive, Idaho Falls, ID 83402; Email, vakanski@uidaho.edu; Phone: (+1)208-757-5422

Abstract

Incorporating human expertise and domain knowledge is particularly important for medical image processing applications, marked with small datasets, and objects of interests in the form of organs or lesions not typically seen in traditional datasets. However, the incorporation of prior knowledge for breast tumor detection is challenging, since shape, boundary, curvature, intensity, or other common medical priors vary significantly across patients and cannot be employed. This work proposes an approach for integrating visual saliency into a deep learning model for breast tumor segmentation in ultrasound images. Visual saliency emphasizes regions that are more likely to attract radiologists’ visual attention and stand out from its surrounding. Our approach is based on a U-Net model and employs attention blocks to introduce visual saliency. Such model forces learning feature representations that prioritize spatial regions with high levels of saliency. The approach is validated using a dataset of 510 breast ultrasound images.

Keywords: Breast ultrasound, medical image segmentation, visual saliency, domain-knowledge enriched learning

Introduction

Computer-aided image analysis can assist radiologists’ interpretation and diagnosis, and reduce error rates as well as the level of stress regarding erroneous diagnosis (Cheng et al. 2010; Inoue et al. 2017; Jalalian et al. 2017; Moon et al. 2011; Wu et al. 2019). For instance, 3 to 6% of all radiologists’ image interpretations contain clinically important errors, and also, a significant variability in the inter- and intra-observer image interpretation is often reported (Elmore et al. 1994; Langlotz et al. 2019; Waite et al. 2016).

The emphasis in this work is on automated computer-aided detection of tumors in breast ultrasound (BUS) images (Xian et al. 2018b). Ultrasound is an important imaging modality for breast cancer detection due to the noninvasive and nonradioactive imaging procedure (e.g., when compared to mammography), high sensitivity, effectiveness for analysis of dense breasts, and also portability of the equipment and relatively lower cost in comparison to other related imaging methods (Guo et al. 2018). On the other hand, interpretation of BUS images is more dependent on radiologists’ skills and expertise because of the low contrast and high level of speckle noise and artifacts in images. Prior research put emphasis on several pertinent tasks, including automated lesion localization, segmentation, and classification (Inoue et al. 2017; Jalalian et al. 2017; Litjens et al. 2017; Moon et al. 2011; Wu et al. 2019). Breast tumor segmentation entails extracting the tumor region in BUS images by demarcating the boundary of the tumor from the background tissue. Based on the segmentation results, subsequent steps typically involve quantitative analysis of the segmented regions and tumor classification. A large volume of research employed various conventional approaches for tumor segmentation using thresholding, region-growing, graph-based approaches, and deformable models (Cai and Wang 2013; Gómez-Flores and Ruiz-Ortega 2016; Huang et al. 2017; Liu et al. 2010; Rodrigues et al. 2015; Xiao et al. 2002). In recent times, deep learning-based models outperformed conventional methods for image segmentation in general, and breast tumor segmentation in particular, and have been the preferred means for addressing this task (Akselrod-Ballin et al. 2017; Chiang et al. 2019; Jaeger et al. 2018; Kooi et al. 2017; Litjens et al. 2017; Yap et al. 2018). In spite of the progress, breast tumor segmentation remains a challenging task, due to the low contrast and presence of noise in images. In addition, the wide-ranging variations in tumors appearance in images exacerbate the development of efficient computational algorithms with strong generalization abilities. Consequently, existing methods lack robustness and consistency when processing images taken with different imaging equipment, where the variations in image intensity, contrast, and density often result in a degraded performance of models that otherwise perform well on custom-built datasets. The lack of large annotated datasets of BUS images and the
An important way to improve the performance of data-driven models is incorporating prior domain-specific knowledge (Nosrati and Hamarneh 2016). On the other hand, incorporating prior knowledge in deep models for breast cancer detection is challenging, because unlike other medical organs—such as the kidney or the heart, whose features naturally lend themselves to the application of shape or boundary priors—breast tumors have a large variability in shape and boundaries from case to case. Extracting other priors in the form of curvature, texture, intensity, or number of regions for breast tumors is also not an option. Our proposed approach incorporates topological and anatomical prior information into a deep learning model for image segmentation. More specifically, maps of visual saliency are employed for integrating image topology knowledge (Xu et al. 2016; Xu et al. 2018). The employed model for visual saliency estimation is formulated as a quadratic optimization problem and is based on calculations of neuron-connectedness between regions in the image (Xian et al. 2016; Xian 2017). The model integrates anatomical prior knowledge by decomposing the tissue layers in images into several layers based on the different appearances of skin, fat, mammary, and muscle layers (Xu et al. 2019). Since breast tumors typically occur in the mammary layer, the model applies higher weights to salient regions belonging to the mammary layer with high probabilities.

In this paper, we integrate the prior knowledge in a deep network by using the attention mechanism (Simonyan et al. 2013), in the form of a pyramid of visual saliency maps. A U-Net architecture is selected as the model for incorporating the priors (Ronneberger et al. 2015). Attention blocks are integrated with the layers of the encoder to force the network to learn feature representations that place spatial attention to target regions with high saliency values. Unlike similar deep learning models that introduce attention blocks by merging internal feature representations from different layers (Chen et al. 2016; Jetley et al. 2018; Oktay et al. 2018b), the proposed approach employs external auxiliary inputs in the form of visual saliency maps for training the model parameters. The main contributions of this paper are: (1) attention enriched deep learning model for integrating prior knowledge of tumor saliency; and (2) confidence level calculation for visual saliency maps.

The paper is organized as follows. The ensuing subsection provides an overview of related works in the literature. The Materials section describes the used image dataset. The proposed approach is presented next in the Methods section, which encompasses the description of the proposed network architecture, attention block, and generation of visual saliency maps. The experimental validation of the approach are provided in the Results section. The Discussion section describes the findings of this work, and the Conclusion section concludes the major contributions of this work.

**Related Works**

Computer-aided segmentation in medical imaging has been an important research topic for several decades, and it encompasses a vast body of work in the published literature. Recent advances in deep learning models (Goodfellow et al. 2016; LeCun et al. 2015) demonstrated great improvements in semantic image segmentation (Badrinarayanan et al. 2017; Chen et al. 2018a; Chen et al. 2018b; He et al. 2015; Lin et al. 2017; Long et al. 2015; Ronneberger et al. 2015; Zhao et al. 2017). Consequently, significant efforts have been devoted toward the implementation and design of deep neural networks for a wide range of medical applications, including segmentation of tumors and lesions (e.g., brain tumor (Kamnitsas et al. 2017), skin lesions (González-Díaz 2017), histopathology images (Chen et al. 2017; Graham et al. 2018; Kumar et al. 2017; Lin et al. 2018; Naylor et al. 2019)), and segmentation of organs (e.g., pancreas (Oktay et al. 2018b), lung (Hu et al. 2019), heart (Oktay et al. 2018a), or head and neck anatomy (Zhu et al. 2019)).

Likewise, the implementation of deep models for breast tumor segmentation has spurred interest in the research community in recent years (Xian et al. 2018b). Whereas the most popular image modality for this task have been ultrasound images (Abraham and Khan 2018; Chiang et al. 2019; Huang et al. 2018; Yap et al. 2018) and digital mammography images (Akselrod-Ballin et al. 2017; Dhungel et al. 2015; Jung et al. 2018; Kooi et al. 2017; Moor et al. 2018; Ribi et al. 2017), a body of literature used MRI (Jaeger et al. 2018), and histology images (Lin et al. 2018). U-Net (Ronneberger et al. 2015) and its numerous variants and modifications are the most commonly used architecture for this problem to date. In spite of the progress, this is still an open research topic, due to challenges related to the inherent presence of noise and low contrast of images, sensitivity of current methods to the used image-acquisition method, equipment, and settings, and the lack of large open datasets of annotated images for training purposes.
Priors in medical image segmentation. Incorporating prior task-specific knowledge for medical image segmentation is important for improved model performance (Nosrati and Hamarneh 2016), and it can be crucial in tasks with small datasets of annotated medical images (i.e., most medical tasks at the present time). Prior knowledge can generally be in the form of shape, boundary, curvature, appearance (e.g., intensity, texture), topology (e.g., connectivity), anatomical information / atlas (structure of tissues or organs), user information (seed points or bounding boxes), moments (size, area, volume), distance (between organs and structures), and other forms. Although recent deep learning-based models have caused a leap of performance in image segmentation over conventional methods, incorporating prior knowledge in deep neural networks has proven to be a difficult task, and consequently, has not been widely investigated. Namely, semantic image segmentation using deep networks typically relies on loss functions that optimize the model predictions at a pixel level, without taking into consideration inter-pixel interactions and semantic correlations among regions at the image level. To integrate prior knowledge into segmentation models, several works have proposed custom loss functions that enforce learning feature representations compatible with the priors. For instance, a loss function that penalizes geometric priors (boundary smoothness) and topological priors (containment or exclusion of lumen in epithelium and stroma) was devised for histology gland segmentation (BenTaieb and Hamarneh 2016). Likewise, loss functions in FCNs that encode a shape prior were proposed for kidney segmentation (Ravishankar et al. 2017), cardiac segmentation (Oktay et al. 2018a), and segmentation of star shapes in skin lesions (Mirikharaji and Hamarneh 2018). The disadvantage of this approach is that the related models are task-specific, and cannot be repurposed for segmentation of other objects of interest in medical images. Another line of research introduces a post-processing step with Conditional Random Fields where the segmentation predictions by a deep learning network are improved through assigning class labels to regions with similar topological properties (Chen et al. 2018a; Havaei et al. 2017; Huang et al. 2018). However, these methods increase the processing complexity and computational expense, and have been mostly replaced in recent years with end-to-end training models. Furthermore, a body of work proposed to incorporate shape priors by redesigning the network architecture. For example, Li et al. (Li et al. 2016) employed a FCN with a VGG-16 base model where shape priors are learned by a consecutive concatenation of the original images with the obtained segmentation maps during several iterations of the procedure. Gonzalez-Diaz (González-Díaz 2017) created probability maps based on the knowledge of the patterns of skin lesions (e.g., dots, globules, streaks, or vascular structures) and merged them with extracted feature maps in a ResNet-based architecture. Furthermore, a boundary prior was incorporated in a deep learning model called DCAN that has two subnetworks for learning concurrently shapes and contour boundaries in histology images (Chen et al. 2017). Yet another class of methods utilize generative models for introducing prior knowledge. E.g., in several early pre-FCN image segmentation models, Boltzmann machines networks were employed for learning shape priors (Chen et al. 2013; Eslami et al. 2014). A more recent research uses variational Bayes autoencoders for incorporating prior anatomical knowledge of the brain geometry in segmentation of MRI images (Dalca et al. 2018).

Despite the potential demonstrated by the above-described research work, to the best of our knowledge, there is no previous studies on the incorporation of prior knowledge in deep models for breast cancer detection. The challenge stems from the fact that unlike other medical organs (e.g., kidney, heart) where shape or boundary priors can be applied, such constraints are not applicable to breast cancer detection, due to the wide difference in the geometry of breast tumors. Analogously, it is difficult to extract generalized prior knowledge regarding curvature, moments, appearance, intensity, or number of regions for breast tumors. In this work we introduce prior topology information in a deep learning segmentation model in the form of region connectivity and saliency. Such prior information is combined with anatomical prior knowledge of the tissue layers in breast images, as explained in the subsequent sections.

Attention mechanism in deep learning. Attention mechanism is an approach in deep networks layer design where the goal is to recognize discriminative features in the inner activation maps and to utilize this knowledge toward enhanced task-specific data representation and improved model performance (Simonyan et al. 2013). This mechanism contributes to suppressing less relevant features and emphasizing more important features for the considered task; e.g., in image classification important features lie in salient spatial locations in the images. Attention mechanism has been integrated in various deep learning models designed for image captioning (Li et al. 2018; Xu et al. 2015), language translation (Bahdanau et al. 2015), and image classification (Jetley et al. 2018;
Wang et al. 2017). In general, attention in deep neural networks is traditionally implemented in two main forms, known as hard and soft attention. The implementation of hard (or stochastic) attention is non-differentiable, the training procedure is based on a sampling technique, and as a consequence, the models are difficult to optimize (Cao et al. 2015; Mnih et al. 2014; Stollenga et al. 2014). Soft (or deterministic) attention models are differentiable and trained with backpropagation; due to these properties they have been the preferred form of implementation (Chen et al. 2016; Jaderberg et al. 2015; Wang et al. 2017). In image processing, the attention mechanism produces a probabilistic map of spatial locations in images, where the parameters of the attention map are learned in end-to-end training. Furthermore, the introduced architecture designs in image processing typically comprise of multiple attention maps with different resolutions, thereby capturing salient features across multiple levels of feature abstraction. For instance, Jetley et al. (Jetley et al. 2018) introduced attention gates at three intermediate layers in a VGG network, and a weighted combination of the attention maps is used in the last layer for image classification. Chen et al. (Chen et al. 2018a) introduce attention blocks in the initial DeepLab model for image segmentation, where attention weights are learned at different scales of a pyramidal feature representation.

Similar, attention gates were introduced in a U-Net architecture (Oktay et al. 2018b) and were employed in medical image processing for segmentation of the pancreas (Oktay et al. 2018b), and for breast tumor and skin lesion segmentation (Abraham and Khan 2018). This type of models uses the extracted features maps in the encoder path of the network for calculation of attention maps, which are afterward merged with the up-sampled features maps in the decoder network, typically via element-wise multiplication. Such design forces the model to encode the locations and shapes of salient regions in extracted representations that are relevant for segmentation of the objects of interest. In the work by Tomita et al. (Tomita et al. 2018) an attention module was implemented in a 3D residual CNN to dynamically identify regions of interest (ROI) in processing high-resolution microscopy images, thus replacing the commonly used approach of sliding window ROI selection, and alleviating the computational burden in processing microscopy images. In a related work to the proposed approach, AttentionNet is designed on top of a ResNeXt encoder-decoder architecture, and applies both spatial and channel attention blocks for segmentation of the anatomical tissue layers in BUS images (Li et al. 2019). Conversely to our method, the authors in (Li et al. 2019) did not apply AttentionNet for breast tumor detection, as well as they used activations maps of the intermediate layers of the network in the attention blocks.

Materials and Methods

The proposed approach is validated on a dataset of 510 breast ultrasound images (Xian et al. 2018a). The dataset is collected from three hospitals: the Second Affiliated Hospital of Harbin Medical University, the Affiliated Hospital of Qingdao University, and the Second Hospital of Hebei Medical University. All images in the dataset are de-identified, and informed consent to the protocol was obtained from all involved patients. Different types of imaging ultrasound devices were employed for acquiring the images, including GE VIVID 7 and LOGIQ E9, Hitachi EUB-6500, Philips iU22, and Siemens ACUSON S2000. This makes the dataset suitable for training data-driven models with enhanced robustness to variations across images from different sources. Image annotation related to the segmentation and delineation of tumors in images was initially performed by three experienced radiologists, followed by voting and creating a single segmentation mask on which all three medical professionals agreed. Afterward, the annotations were reviewed by a senior radiologist expert, who either approved, or if needed, applied corrections and amendments to the segmentation boundaries (Xian et al. 2018a).

Network Architecture

The proposed network is based on the well-known U-Net architecture (Ronneberger et al. 2015), which consists of fully convolutional encoder and decoder sub-networks with skip connections. The layers in the encoder employ a cascade of convolutional and max-pooling layers, which reduce the resolution of input images and extract increasingly abstract features. The decoder comprises of convolutional and up-sampling layers that provide an expanding path for recovering the spatial resolution of extracted feature maps to the initial level of the input images. A unique characteristic of the U-Net architecture is the presence of skip connections from the feature maps in encoder’s contracting path to the corresponding layers in the decoder. The features from the respective encoder’s and decoder’s
layers are merged via concatenation that allows to recover the spatial accuracy of the objects in images, and improves the resulting segmentation masks. Namely, although the central layer of the network offers high-level features with semantic rich data representation and a large receptive field, it also has low level of spatial context detail due to the down-sampling max-pooling layers along the contracting path, and impacts the localization accuracy around the object boundaries in the predictions. The skip connections provide a means to transmit low-level feature information from the initial high-resolution layers in the encoder to the reconstructing layers in the decoder, thereby restoring the local spatial information in predicted segmentations. Despite the introduction of deeper and more powerful models for image segmentation in recent years, the U-Net architecture has remained popular especially in medical image segmentation, where datasets have small size and large models can overfit on the available sets.

A graphical representation of the proposed model is presented in Figure 1. Beside the main input consisting of BUS images, the network has an auxiliary input consisting of the corresponding salient maps. Attention blocks introduce salient maps with reduced scale in all layers on the contracting path of the encoder in a form of an image pyramid. This enforces the network to focus the attention onto regions in the saliency maps with high intensity values. More specifically, the introduced attention blocks put more weights on areas in the extracted feature maps at each layer that have higher levels of saliency in the salient maps. Thus, the topology of the salient maps influences the learned feature representations.

**Figure 1:** Architecture of the proposed U-Net model with salient attention. The model uses BUS images and saliency maps as inputs, and produces segmentation probability maps as outputs.

The input images and saliency maps have dimensions of 256 × 256 pixels, and they have one channel due to the nature of ultrasound images. The number of convolutional filters per layer in the network is (32, 32, 64, 64, 128), which is reduced in comparison to the original U-Net, to account for the relatively small dataset. The output segmentation probability maps have the same spatial dimension as the inputs. The proposed network is trained in an end-to-end fashion; however, the saliency maps are precomputed and used at training and inference.

**Attention Blocks**

A block diagram of the attention block $n$ is depicted in Figure 2. The input feature maps to the attention block are denoted $F_n = \{f_1, f_2, ..., f_{k_n}\}$, where each feature map has horizontal and vertical spatial dimensions $256/2^{(n-1)} \times 256/2^{(n-1)}$ for the block in the layer level $n \in \{1, 2, 3, 4\}$. The symbol $k_n$ is the channel dimension of the feature maps in block $n$, i.e., $k_n \in \{32, 64, 128\}$. For example, the output activations of the convolutional ‘Conv 64’ layer entering Attention block 4 have dimensions of $32 \times 32 \times 64$, i.e., $256/2^3 \times 256/2^3 \times k_4$.

The saliency map $S$ is down-sampled through a max-pooling layer, denoted here $S_n$, matching the spatial dimension of the input feature maps $F_n$ in attention block $n$. Next, 1x1 convolutions followed by ReLU activation functions are used to increase the number of channels of the saliency map to 128. An element-wise sum block performs
addition producing intermediate maps \( I_n \) of size \( 256/2^n \times 256/2^n \times 128 \). The intermediate maps \( I_n \) are further refined through a series of linear \( 128 \times 3 \times 3 \) convolutions and \( 1 \times 1 \times 1 \) convolution, and nonlinear ReLU activations. A sigmoid activation function normalizes the values of the activation maps into the \([0, 1]\) range. The produced output is the attention map \( A = (\alpha_i) \) with a spatial size of \( 256/2^n \times 256/2^n \times 1 \), where the attention coefficients \( \alpha_i \) have scalar values for each pixel \( i \). Next, soft attention is applied via element-wise multiplication of the attention map with the max-pooled features, i.e., \( O_n = A \ast P_n \). The output activation maps \( O_n \) with size \( 256/2^n \times 256/2^n \times k_n \) are further propagated to the next layer.

![Diagram](image)

**Figure 2:** Attention block \( n \), for \( n \in \{1, 2, 3, 4\} \). Inputs to the block are feature maps from layer \( n \) with spatial dimension \( 256/2^{(n-1)} \times 256/2^{(n-1)} \) with \( k_n \) number of channels, and a salient map, and the output are down-sampled weighted maps with spatial dimension \( 256/2^n \times 256/2^n \) and \( k_n \) number of channels.

The design of the attention block was inspired by the attention gates in (Okty et al. 2018b) and (Jetley et al. 2018). Differently from these two works, where the attention blocks employ activation maps from the intermediate layers in the model as saliency maps for enhancing the discriminative characteristics of extracted intermediary features, the proposed attention block in this work utilizes precomputed saliency maps that point out to target spatial regions. If the attention block in this work applies directly the self-attention blocks described in (Abraham and Khan 2018; Jetley et al. 2018; Oktay et al. 2018b) the segmentation performance of the model does not improve. The reason for that lies in the distribution of salient regions in the used maps, since in many images background non-tumor areas have certain level of saliency in the salient maps; consequently, placing equal attention weights on all salient regions leads to higher level of false positive errors and degraded performance. The introduction of additional \( 3 \times 3 \) and \( 1 \times 1 \) convolutional layers for feature refinement in the presented salient attention block was conducive toward improved segmentation outputs, which was confirmed via empirical validation of the proposed layers design.

**Saliency Maps**

Visual saliency estimation is an important paradigm for automatic tumor detection in BUS images, where the aim is to model the level of saliency of image regions in correspondence to the capacity to attract radiologists' visual attention (Shao et al. 2015; Xie et al. 2017). For an input image, the output of such models is a visual saliency map with assigned saliency values in the \([0, 1]\) range to every image pixel. High saliency value indicates a high probability that the pixel belongs to a tumor.

The adopted approach for generating saliency maps of BUS images is based on our previous work (Xian et al. 2016; Xian 2017; Xu et al. 2016; Xu et al. 2018; Xu et al. 2019). In particular, the task of visual saliency estimation
is formulated as a quadratic programming optimization that integrates high-level image information and low-level saliency assumptions. Our most recent work on this topic (Xu et al., 2019) introduces additional constraints in the model related to the breast anatomy. The anatomical information of breast tissue layers, different appearances in skin, fat, mammary and muscle layers, is utilized to decompose the image into several layers through the Neutro-Connectedness (NC) approach (Xian, 2017; Xian et al., 2016). The anatomical layers information and the fact that tumors are present predominantly in the mammary layer are used in our framework as high-level image information for saliency estimation. Two low-level saliency assumptions are utilized in the framework as well: 1) adaptive-center bias assumption forces the regions nearer the adaptive center to have higher saliency values; 2) the region-correlation assumption forces the similar regions to have similar saliency values. The extensive experiments in (Xu et al. 2019) showed the new model with anatomical knowledge generated much better performance than the other models on the dataset (Xian, 2018b). Full implementation details can be found in the respective publications.

Examples of breast images and corresponding saliency maps are presented in Figure 3. The top row in the figure shows five BUS images, and the middle row displays the ground truth segmentation masks provided by radiologists. The bottom row displays the saliency maps for the images. One can note that the saliency maps assign a value to every pixel regarding the probability of belonging to a tumor, and differently from the ground truth masks, saliency values are assigned to background regions in images as well. Furthermore, the saliency maps are generated in an unsupervised manner, i.e., the information of the ground truth is not used by the saliency estimation model.

The incorporation of saliency maps into a deep learning model as a complementary prior information is based on an assumption that the areas in images with high saliency values correspond to high probability of tumor presence. Therefore, it is important that the saliency maps are of adequate quality and provide reliable information regarding the tumor locations. Otherwise, poor quality saliency maps can degrade the model performance.

![Figure 3: Examples of saliency maps with varying level of quality. Top row: original BUS image; Middle row: ground truth mask; Bottom row: saliency map.](image)

The selected five examples of saliency maps depicted in Figure 3 have different levels of quality. More specifically, a map is considered of satisfactory quality when the location and intensity of the tumor region is clearly discernable in the saliency map. The example in the middle column in Figure 3 with moderate quality indicates the tumor location correctly, but the tumor shape and boundary do not match very well the ground truth, which may cause errors in the edge segmentation when applied to a deep network. For the case with low quality in Figure 3 there are several regions with similar area and saliency values, and it is not clear which of these regions may be tumors. Lastly,
the saliency map with poor quality in Figure 3 assigns zero saliency values to the tumor region, and completely misses the tumor.

In order to account for the cases with lower quality of saliency maps, we employed an algorithm that calculates the level of confidence in the saliency maps, and subsequently, eliminates the maps with low confidence. The approach is based on the following parameters: contour area $A_c = \sum_j p_j$ is the number of pixels of a contour $c$ in an image with a saliency value per pixel $p_j$ greater than a threshold value; cumulative intensity $I_c = \sum_j S(p_j)$ calculates the sum of the saliency values for the pixels in contour $c$; and mean intensity $M_c = \sum_j S(p_j)/A_c$ of a contour $c$ is calculated as the ratio of the cumulative intensity and the area. The first rule in Algorithm 1 states that if the contour with the largest cumulative intensity $I_{arg\max(I_{1:c})}$ has similar cumulative intensity to the second largest contour and its mean intensity $M_{arg\max(I_{1:c})}$ is not the highest of all contours (see Figure 4, left column), then eliminate the saliency map from the set. The second rule is similar to the first rule, and takes into account cases with larger ambiguities in the cumulative intensity and mean intensity of contours (Figure 4, middle column). The third rule considers the cases when a contour has high mean saliency intensity but smaller cumulative intensity than other contours in the image (see Figure 3, right column). The parameters in the algorithm are empirically set to: $a_1 = 2$, $a_2 = 3$, $a_3 = 0.2$, and $a_4 = 0.55$. In total 52 saliency maps satisfied the given conditions and were removed from the original set of 562 images, resulting in a reduced set of 510 images. That is, approximately 91% of the saliency maps are with high level of confidence. Having a low level of confidence for a saliency map does not necessarily mean that the saliency map is not correct: e.g., one can argue that the saliency for the example in the middle column in Figure 4 is correct. Rather, the proposed algorithm is designed to identify saliency maps with ambiguities regarding the spatial regions for tumor existence. The algorithm takes as inputs only the saliency maps, and it does not use the knowledge of the ground truth in estimating the level of confidence.

**Algorithm 1:** Confidence level calculation for saliency maps

For saliency map $i = 1$: 

For contour $c = 1$: $C$

if $I_{arg\max(I_{1:c})} < a_1 I_{arg\max(I_{1:c})-1}$ and $M_{arg\max(I_{1:c})} < M_{arg\max(M_{1:c})}$

or if $I_{arg\max(I_{1:c})} < a_2 I_{arg\max(I_{1:c})-1}$ and $M_{arg\max(I_{1:c})} + a_3 < M_{arg\max(M_{1:c})}$

or if $M_{arg\max(M_{1:c})} > a_4$ and $arg\max(M_{1:c}) \neq arg\max(I_{1:c})$

Remove saliency map $i$ from the set

**Results**

**Evaluation Metrics**

We use Dice score coefficient (DSC), Jaccard index (JI), true positives ratio (TPR), false positives ratio (FPR), and global accuracy (ACC) to evaluate the model performance:

\[
DSC = \frac{2|A_g \cap A_p|}{|A_g| + |A_p|} \tag{1}
\]

\[
JI = \frac{|A_g \cap A_p|}{|A_g \cup A_p|} \tag{2}
\]

\[
TPR = \frac{|A_g \cap A_p|}{|A_p|} \tag{3}
\]

\[
FPR = \frac{|A_g \cup A_p - A_g|}{|A_g|} \tag{4}
\]

\[
ACC = \frac{|A_g \cap A_p| + |A_g - A_g \cup A_p|}{|A_g| + |A_p|} \tag{5}
\]
In the above equations, $A_g$ is the set of pixels that belong to a tumor region in the ground truth segmented images, $\overline{A_g}$ is the set of pixels that belong to the background region without tumors in the ground truth segmented images, and $A_p$ is the corresponding set of pixels that are predicted to belong to a tumor region by the segmentation method. It is important to note that FPR is calculated as the ratio divided by the number of positives (i.e., pixels in tumor regions in the ground truth masks), as opposed to a ratio divided by the number of negatives (i.e., pixels in the background regions in the ground truth masks) as it is often defined in related tasks. Since the positive regions are smaller in BUS tumor segmentation, the selected formulation for FPR is more descriptive for this task. Whereas the TPR and FPR calculate specific aspects of the predicted segmentation, they are not comprehensive in the sense that high value of TPR alone does not necessarily warrant satisfactory performance. The DSC and JI metrics have abilities to capture the method performance independently, and are most commonly used metrics for evaluation of image segmentation. Global accuracy is less commonly reported in image segmentation.

![Figure 4: Examples of eliminated saliency maps from the original dataset. Top row: original BUS image; Middle row: ground truth mask; Bottom row: saliency map.](image)

Another metric that is often used for performance evaluation is area under the curve (AUC) of ROC (receiver operating characteristic) (i.e., AUC-ROC score). The ROC curve is obtained by plotting the TPR on the vertical axis and FPR on the horizontal axis, i.e., it combines the impact of both TPR and FPR on the methods’ performance. The area under the curve is commonly reported and used for comparative analysis. For all the above metrics, the values are in the $[0, 1]$ range, where higher values indicate improved performance (except for FPR where low values are preferred).

**Implementation Details**

The proposed approach is validated on the described dataset of BUS images. We used five-fold cross-validation, i.e., four folds (80% of images) are used for training, and one fold (20% of images) is used for testing. Validation during training is performed on 20% of the training set of images. All images in the dataset are first resized to a 256 x 256 pixels resolution. Since we focus on understanding the impact of the introduced salient attention on the model performance, we did not apply image augmentation.

The proposed model is trained with randomly initialized weights using Xavier normal initialization (Glorot and Bengio 2010). Dice loss function was used for training, defined as

$$\mathcal{L} = 1 - DSC = 1 - \frac{2|A_g \cap A_p|}{|A_g| + |A_p|}$$

(6)
where the same notation is preserved, i.e., $A_g$ and $A_p$ denote the ground truth and predicted masks, respectively.

The models were implemented using TensorFlow and Keras libraries. The network was trained by using adaptive moment estimation optimizer (Adam) with a learning rate of $10^{-4}$, and a batch size of 4 images. The training was stopped when the loss of the validation set did not improve for 20 epochs.

**Evaluation and Comparative Analysis**

The experimental validation of the proposed approach is based on a comparative analysis of the following three models:

(1) U-Net. It is the basic U-Net implementation;
(2) U-Net-SA. It applied the proposed salient attention approach; and
(3) U-Net-SA-C. It is a model with salient attention applied to a modified version where only one contour with the highest saliency is extracted in each salient map.

Examples of input BUS images, ground truth masks, saliency maps, and output segmentation maps by the models are presented in Figure 5. The values of the performance metrics are provided in Table 1. For the BUS images displayed in Figure 5, the segmentation outputs by the U-Net model are inferior in comparison to the predicted masks produced by the models with salient attention U-Net-SA and U-Net-SA-C. One particular aspect of improved performance entails the false positive predictions by U-Net (see rows A-G in Figure 5). In these cases, U-Net produces positive predictions of tumor presence for image regions that don’t belong to a tumor. The attention models U-Net-SA and U-Net-SA-C benefited from the information in the salient maps, which led to a reduced rate of false positive predictions in A-G. This is especially noticeable in rows B, E, and G that have high quality salient maps, resulting in great improvement over the prediction by the basic U-Net model. Furthermore, improved performance with respect to the true positive predictions by U-Net is displayed for rows H and I in Figure 5.

**Table 1:** Performance evaluation metrics for models without and with salient attention. The shown values correspond to the average and standard deviation (in parenthesis) per fold in a 5-fold cross-validation.

| Model            | Dataset                        | DSC   | JI (IOU)  | TPR   | FPR   | ACC   | AUC-ROC |
|------------------|--------------------------------|-------|-----------|-------|-------|-------|---------|
| U-Net            | Basic U-Net                    | 0.894 | 0.821     | 0.903 | 0.107 | 0.978 | 0.951   | (±0.013) | (±0.017) | (±0.011) | (±0.019) | (±0.002) | (±0.006) |
| U-Net-SA         | Salient attention              | 0.901 | 0.832     | 0.904 | 0.092 | 0.979 | 0.955   | (±0.013) | (±0.014) | (±0.016) | (±0.008) | (±0.001) | (±0.002) |
| U-Net-SA-C       | Salient attention, one contour/image | **0.905** | **0.838** | **0.910** | **0.089** | **0.980** | **0.957** | (±0.013) | (±0.014) | (±0.011) | (±0.012) | (±0.001) | (±0.004) |

The provision of salient maps for these two cases helps the model to focus on target regions with high saliency, leading to higher true positives rate of the segmentation masks by U-Net-SA over the basic U-Net model. In addition, rows J and K provide examples where the geometry of the salient regions in the saliency maps contributes to more accurate predictions of the proposed models in comparison to U-Net. Cases C and I are instances of BUS images with small size tumors, where the salient attention models successfully located the tumor regions. As explained earlier, the U-Net-SA-C model employs salient maps with one contour with the highest saliency intensity, and in many images it further improves the segmentation outputs. This is noticeable in row A in Figure 5, where the false positives in the segmentation are reduced in comparison to U-Net-SA. However, U-Net-SA-C model is based on an assumption that there is only one tumor in the images, which may not always be the case.

The results in Table 1 show the average and standard deviation (in parenthesis) per fold in the 5-fold cross-validation procedure for the three deep models. The obtained values indicate that the models with salient attention U-Net-SA and U-Net-SA-C outperform the basic U-Net network without attention blocks for all performance metrics.
The model U-Net-SA-C trained on the dataset with a single contour in the salient maps produced improved segmentation performance in comparison to U-Net-SA.

Next, a comparison of our salient attention model for tumor segmentation U-Net-SA and three respective deep models for image segmentation is provided in Table 2. The dataset with 510 images is used for training the models. For a fair comparison, all models are trained in the same manner as the proposed architecture, i.e., 5-fold cross-validation, batch size of 4, Xavier normal weights initialization, dice loss, Adam optimizer, and a stopping criterion of 20 epochs of non-improved validation loss. Due to the relatively small size of the dataset, for the comparison we selected smaller versions of the models. For instance, DenseNet is based on a network with 26 layers, and for PSPNet (that requires a base model) the small residual model ResNet18 is employed. The learning rate is fine-tuned for the different models, where an initial learning rate is selected, and when the validation loss does not improve for 10 epochs, the learning rate is reduced by a certain step size. The procedure is repeated until a preset value for the learning rate is reached. The details regarding the used learning rates for the different models are provided in Table 2. Our proposed U-Net-SA model listed last in the table outperformed the other deep learning networks for image segmentation on most of the employed performance metrics.

Table 2: Values of the performance metrics for tumor segmentation by different models. The shown values correspond to the average and standard deviation (in parenthesis) per fold in a 5-fold cross-validation. LR represents the used learning rates for training the models.

| Model          | Training Setting                      | DSC (±0.011) | JI (IOU) (±0.015) | TPR (±0.019)  | FPR (±0.014)  | ACC (±0.002) | AUC-ROC (±0.004) |
|----------------|--------------------------------------|--------------|------------------|--------------|--------------|--------------|------------------|
| Seg-Net        | LR=8·10⁻⁴, decreased by 0.5 after 10 epochs until 1·10⁻⁴ | 0.889        | 0.811            | 0.877        | 0.088        | 0.977        | 0.957            |
| DenseNet-26    | LR=1·10⁻³, decreased by 0.1 after 10 epochs until 1·10⁻⁴ | 0.888        | 0.818            | 0.886        | 0.093        | 0.978        | 0.958            |
| PSPNet-ResNet18| LR=1·10⁻⁴, decreased by 0.5 after 10 epochs until 5·10⁻⁴, images size of 384x384 pix. | 0.886        | 0.808            | 0.884        | 0.107        | 0.976        | 0.953            |
| Ours: U-Net-SA | LR = 1·10⁻⁴                           | 0.901        | 0.832            | 0.904        | 0.092        | 0.979        | 0.955            |

Discussion

Based on the evaluation results presented in Table 1, the models with attention blocks outperformed the basic U-Net model. In addition, if only one contour with the highest saliency is extracted in the salient maps (the U-Net-SA-C model), the performance improves further. This can be explained by the increasing spatial attention to a single salient region in the maps, resulting in reduced false positives in the outputs. As we mentioned earlier, this is based on an assumption that there is only one tumor in the images, which may not always be the case.

The design of the attention blocks has an impact on the segmentation output; therefore, we investigated several alternatives for the block layers and their parameters. Compared to similar attention blocks in deep models (Chen et al. 2016; Jetley et al. 2018; Oktay et al. 2018b), the used block in this work requires additional feature refinement by using convolutional 3x3 and 1x1 layers. The refinement layers balance the impact of inaccurate boundaries of the regions in salient maps on the learned features. In other words, the saliency maps do not provide accurate local information of the edges and boundaries of tumors in images, but rather, they provide global information of the spatial probability regarding the presence of tumors. Larger values of the attention coefficients put more emphasis on the edges and boundaries in salient maps and can reduce the segmentation outputs. The use of additional refinement layers lessens the values of the attention coefficients, and results in improved tumor segmentation.
Figure 5: Segmentation results. First column: original BUS image; Second column: ground truth mask; Third column: saliency map; Fourth column: segmentation mask produced by U-Net; Fifth column: segmentation mask produced by U-Net-SA; Sixth column: segmentation mask produced by U-Net-SA-C.
One limitation of the presented approach is that it relies on the quality of saliency maps. Using low quality maps can at best not improve the results, or result in degraded performance. To deal with this shortcoming, we proposed an algorithm that calculates a confidence score and eliminates the saliency maps with low confidence in their level of quality. Whereas visual saliency estimation is not the focus of this work, improvements in the models for visual saliency estimation can lead to improved segmentation by the proposed approach.

Avenues for future work include investigation of custom loss functions in deep learning models for encoding prior information, and working with medical partners to obtain annotated images with breast tissue layers and afterward integrating such anatomical prior with salient maps in a unified segmentation model.

Conclusion
This paper proposes a novel deep learning architecture that incorporates radiologists’ visual attention for breast tumor segmentation. The proposed architecture consists of a variant of the basic U-Net model with attention blocks integrated along the contracting path in the layers of the encoder. The proposed attention blocks allow the deep learning model to suppress spatial regions with low saliency values, and respectively, to focus on regions with high saliency values. The attention blocks use multi-scaled versions of the saliency maps. The approach is validated on a dataset of 510 images, and the results demonstrate improved segmentation performance. The importance of this work stems from the difficulties in incorporating priors into deep learning models for medical image processing, and in particular for segmentation of breast ultrasound images, where most of the traditionally used prior forms cannot be applied.

Acknowledgments
This work was supported by the Center for Modeling Complex Interactions (CMCI) at the University of Idaho through NIH Award #P20GM104420. We would like to thank Fei Xu for providing the visual saliency maps from her latest research and for her constructive feedback and review of the paper.

References
Abraham N, Khan NM. A Novel Focal Tversky Loss Function With Improved Attention U-Net for Lesion Segmentation. 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019) 2018, pp. 683–687.
Akseplod-Ballin A, Karlinsky L, Hazan A, Bakalo R, Horesh AB, Shoshan Y, Barkan E. Deep Learning for Automatic Detection of Abnormal Findings in Breast Mammogram. Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support - Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017. Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, 2017, Proceedings 2017. pp. 321–329.
Badrinarayanan V, Kendall A, Cipolla R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. IEEE Trans Pattern Anal Mach Intell 2017;39:2481–2495.
Bahdanau D, Cho K, Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate. In: Bengio Y, LeCun Y, eds. 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings 2015 [cited 2019 Jul 24].
BenTaieb A, Hamarneh G. Topology Aware Fully Convolutional Networks for Histology Gland Segmentation. In: Ourselin S, Joskowicz L, Sabuncu MR, Únal GB, Wells W, eds. Medical Image Computing and Computer-Assisted Intervention - MICCAI 2016 - 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 2016. pp. 460–468.
Cai L, Wang Y. A phase-based active contour model for segmentation of breast ultrasound images. 6th International Conference on Biomedical Engineering and Informatics, BMEI 2013, Hangzhou, China, December 16-18, 2013 2013. pp. 91–95.
Cao C, Liu X, Yang Y, Yu Y, Wang J, Wang Z, Huang Y, Wang L, Huang C, Xu W, Ramanan D, Huang TS. Look and Think Twice: Capturing Top-Down Visual Attention with Feedback Convolutional Neural Networks. 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015 IEEE Computer Society, 2015. pp. 2956–2964.
Chen F, Yu H, Hu R, Zeng X. Deep Learning Shape Priors for Object Segmentation. 2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, June 23-28, 2013 IEEE Computer Society, 2013. pp. 1870–1877.
Chen H, Qi X, Yu L, Dou Q, Qin J, Heng P-A. DCAN: Deep contour-aware networks for object instance segmentation from histology images. Medical Image Analysis 2017;36:135–146.
Chen L-C, Papandreou G, Kokkinos I, Murphy K, Yuille AL. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. IEEE Trans Pattern Anal Mach Intell 2018a;40:834–848.
Chen L-C, Yang Y, Wang J, Xu W, Yuille AL. Attention to Scale: Scale-Aware Semantic Image Segmentation. 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016 IEEE Computer Society, 2016. pp. 3640–3649.
Langlotz CP, Allen B, Erickson BJ, Kalpathy-Cramer J, Bigelow K, Cook TS, Flanders AE, Lungren MP, Mendelson DS, Rudie JD, Wang G, Kandarpa K. A Roadmap for Foundational Research on Artificial Intelligence in Medical Imaging: From the 2018 NIH/RSNA/ACR/The Academy Workshop. Radiology 2019;291:781–791.

LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436–444.

Li H, Cheng J-Z, Chou Y-H, Qin J, Huang S, Lei B. AttentionNet: Learning Where to Focus via Attention Mechanism for Anatomical Segmentation of Whole Breast Ultrasound Images. 16th IEEE International Symposium on Biomedical Imaging, ISBI 2019, Venice, Italy, April 8-11, 2019 IEEE, 2019. pp. 1078–1081.

Li K, Harirhan B, Malik J. Iterative Instance Segmentation. 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016 IEEE Computer Society, 2016. pp. 3659–3667.

Li L, Tang S, Zhang Y, Deng L, Tian Q. GLA: Global-Local Attention for Image Description. IEEE Transactions on Multimedia 2018;20:726–737.

Lin G, Milan A, Shen C, Reid ID. RefineNet: Multi-path Refinement Networks for High-Resolution Semantic Segmentation. 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017 IEEE Computer Society, 2017. pp. 5168–5177.

Lin H, Chen H, Dou Q, Wang L, Qin J, Heng P-A. ScanNet: A Fast and Dense Scanning Framework for Metastatic Breast Cancer Detection from Whole-Slide Image. 2018 IEEE Winter Conference on Applications of Computer Vision, WACV 2018, Lake Tahoe, NV, USA, March 12-15, 2018 IEEE Computer Society, 2018. pp. 539–546.

Litjens GJS, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafourian M, Laak JAWM van der, Ginneken B van, Sánchez CI. A survey on deep learning in medical image analysis. Medical Image Analysis 2017;22:60–88.

Liu B, Chen H-D, Huang J, Tian J, Tang X, Liu J. Fully automatic and segmentation-robust classification of breast tumors based on local texture analysis of ultrasound images: Pattern Recognition 2010;43:280–298.

Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7–12, 2015 IEEE Computer Society, 2015. pp. 3431–3440.

Mirikharaji Z, Harmanneh G. Star Shape Prior in Fully Convolutional Networks for Skin Lesion Segmentation. In: Frangi AF, Schnabel JA, Davatzikos C, Alberola-López C, Fichtinger G, eds. Medical Image Computing and Computer Assisted Intervention - MICCAI 2018 - 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part IV Springer, 2018. pp. 737–745.

Mnih V, Heess N, Graves A, Kavukcuoglu K. Recurrent Models of Visual Attention. In: Ghahramani Z, Welling M, Cortes C, Lawrence ND, Weinberger KQ, eds. Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada 2014 [cited 2019 Jul 24]. pp. 2204–2212.

Moon WK, Shen Y-W, Huang C-S, Chiang L-R, Chang R-F. Computer-aided diagnosis for the classification of breast masses in automated whole breast ultrasound images. Ultrasound Med Biol 2011;37:539–548.

Moor T de, Rodríguez R, Braz R, Pereira M, Moutinho J, Pinheiro AMG. A Two Step Segmentation Method for Breast Ultrasound Masses Based on Multi-resolution Analysis. Ultrasound in Medicine & Biology 2015;41:1737–1748.

Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab N, Hornegger J, Ill WMW, Frangi AF, eds. Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015 - 18th International Conference Munich, Germany, October 5 - 9, 2015, Proceedings, Part III Springer, 2015. pp. 234–241.

Shao H, Zhang Y, Xian M, Cheng H-D, Xu F, Ding J. A saliency model for automated tumor detection in breast ultrasound images. 2015 IEEE International Conference on Image Processing, ICIP 2015, Quebec City, QC, Canada, September 27-30, 2015 2015. pp. 1424–1428.

Simonyan K, Vedaldi A, Zisserman A. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. arXiv:1312.6034 [cs] 2013 [cited 2019 Jul 24].

Stollenga MF, Masci J, Gomez FJ, Schniiderhuber J. Deep Networks with Internal Selective Attention through Feedback Connections. In: Ghahramani Z, Welling M, Cortes C, Lawrence ND, Weinberger KQ, eds. Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada 2014 [cited 2019 Jul 29]. pp. 3545–3553.
Tomita N, Abdollahi B, Wei J, Ren B, Suriawinata AA, Hassanpour S. Finding a Needle in the Haystack: Attention-Based Classification of High Resolution Microscopy Images. CoRR 2018 [cited 2019 Jul 25];abs/1811.08513.

Waite S, Scott J, Gale B, Fuchs T, Kolla S, Reede D. Interpretive Error in Radiology. American Journal of Roentgenology 2016;208:739–749.

Wang F, Jiang M, Qian C, Yang S, Li C, Zhang H, Wang X, Tang X. Residual Attention Network for Image Classification. 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017 IEEE Computer Society, 2017. pp. 6450–6458.

Wu N, Phang J, Park J, Shen Y, Huang Z, Zorin M, Jastrzebski S, et al. Deep Neural Networks Improve Radiologists’ Performance in Breast Cancer Screening. CoRR 2019;abs/1903.08297.

Xian M. Neutro-Connectedness Theory, Algorithms and Applications. PhD Thesis 2017;101.

Xian M, Zhang Y, Cheng H, Xu F, Ding J. Neutro-Connectedness Cut. IEEE Transactions on Image Processing 2016;25:4691–4703.

Xian M, Zhang Y, Cheng H-D, Xu F, Huang K, Zhang B, Ding J, Ning C, Wang Y. A Benchmark for Breast Ultrasound Image Segmentation (BUSIS). CoRR 2018a;abs/1801.03182.

Xian M, Zhang Y, Cheng H-D, Xu F, Zhang B, Ding J. Automatic breast ultrasound image segmentation: A survey. Pattern Recognition 2018b;79:340–355.

Xiao G, Brady M, Noble JA, Zhang Y. Segmentation of Ultrasound B-mode Images with Intensity Inhomogeneity Correction. IEEE Trans Med Imaging 2002;21:48–57.

Xie Y, Chen K, Lin J. An Automatic Localization Algorithm for Ultrasound Breast Tumors Based on Human Visual Mechanism. Sensors 2017;17:1101.

Xu F, Xian M, Cheng H-D, Ding J, Zhang Y. Unsupervised saliency estimation based on robust hypotheses. 2016 IEEE Winter Conference on Applications of Computer Vision, WACV 2016, Lake Placid, NY, USA, March 7-10, 2016 2016. pp. 1–6.

Xu F, Xian M, Zhang Y, Huang K, Cheng H-D, Zhang B, Ding J, Ning C, Wang Y. A Hybrid Framework for Tumor Saliency Estimation. 24th International Conference on Pattern Recognition, ICPR 2018, Beijing, China, August 20-24, 2018 2018. pp. 3935–3940.

Xu F, Zhang Y, Xian M, Cheng H-D, Zhang B, Ding J, Ning C, Wang Y. Tumor Saliency Estimation for Breast Ultrasound Images via Breast Anatomy Modeling. CoRR 2019;abs/1906.07760.

Xu K, Ba JL, Kiros R, Cho K, Courville A, Salakhutdinov R, Zemel RS, Bengio Y. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37 JMLR.org, 2015 [cited 2019 Jul 24]. pp. 2048–2057.

Yap MH, Pons G, Marti J, Ganau S, Sentí M, Zwiggelaar R, Davison AK, Martí R. Automated Breast Ultrasound Lesions Detection Using Convolutional Neural Networks. IEEE J Biomedical and Health Informatics 2018;22:1218–1226.

Zhao H, Shi J, Qi X, Wang X, Jia J. Pyramid Scene Parsing Network. 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017 IEEE Computer Society, 2017. pp. 6230–6239.

Zhu W, Huang Y, Zeng L, Chen X, Liu Y, Qian Z, Du N, Fan W, Xie X. AnatomyNet: Deep learning for fast and fully automated whole-volume segmentation of head and neck anatomy. Medical physics 2019;46:576–589.