Research Article

Weakly- and Semisupervised Probabilistic Segmentation and Quantification of Reverberation Artifacts

Alex Ling Yu Hung,1 Edward Chen,2 and John Galeotti1,2

1Department of Biomedical Engineering, Carnegie Mellon University, Pittsburgh, PA, USA
2Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA

Correspondence should be addressed to Alex Ling Yu Hung; lingyuh@andrew.cmu.edu

Received 27 July 2021; Accepted 7 February 2022; Published 1 March 2022

Copyright © 2022 Alex Ling Yu Hung et al. Exclusive Licensee Suzhou Institute of Biomedical Engineering and Technology, CAS. Distributed under a Creative Commons Attribution License (CC BY 4.0).

Objective and Impact Statement. We propose a weakly- and semisupervised, probabilistic needle-and-reverberation-artifact segmentation algorithm to separate the desired tissue-based pixel values from the superimposed artifacts. Our method models the intensity decay of artifact intensities and is designed to minimize the human labeling error.

Introduction. Ultrasound image quality has continually been improving. However, when needles or other metallic objects are operating inside the tissue, the resulting reverberation artifacts can severely corrupt the surrounding image quality. Such effects are challenging for existing computer vision algorithms for medical image analysis. Needle reverberation artifacts can be hard to identify at times and affect various pixel values to different degrees. The boundaries of such artifacts are ambiguous, leading to disagreement among human experts labeling the artifacts. Methods. Our learning-based framework consists of three parts. The first part is a probabilistic segmentation network to generate the soft labels based on the human labels. These soft labels are input into the second part which is the transform function, where the training labels for the third part are generated. The third part outputs the final masks which quantifies the reverberation artifacts.

Results. We demonstrate the applicability of the approach and compare it against other segmentation algorithms. Our method is capable of both differentiating between the reverberations from artifact-free patches and modeling the intensity fall-off in the artifacts.

Conclusion. Our method matches state-of-the-art artifact segmentation performance and sets a new standard in estimating the per-pixel contributions of artifact vs underlying anatomy, especially in the immediately adjacent regions between reverberation lines. Our algorithm is also able to improve the performance of downstream image analysis algorithms.

1. Introduction

Ultrasound imaging is low-cost and safe. In addition, its real-time operation is perfect for monitoring needle insertions and other clinical interventions. However, highly reflective parallel surfaces, such as needle walls, can create significant reverberation artifacts because the sound wave reverberates between the posterior and anterior surfaces of the object [1]. When the amount of reflected energy is significant, it manifests itself as an additional echo from the same surface. The reverberation artifacts are relatively bright, looking like actual boundaries which sometimes would overlap with tissue present in the image. Such artifacts not only can be caused by needles and other metallic objects but they also might be the result of certain anatomical structures with large acoustic impedance [2]. This kind of artifact can cloud clinicians’ judgement [3] and confuse medical image analysis algorithms [4, 5]. For some pixels, it can be difficult to differentiate whether the pixel is an artifact, or to assign a percentage to the pixel indicating how much of the pixel’s value is artifact or actual tissue measurement. The brightness of artifacts somewhat predictably falls off as they get further away from the reflective object, but the artifacts have uncertain boundaries and differing intensity distributions. Consequently, pixel-wise labeling is challenging and time consuming for annotators, who may have considerable differences in their annotations. An example is shown in Figure 1(a), where different annotators agree on the general location of the reverberation artifacts but they disagree on the details. Besides, it can also be seen that it is extremely hard for annotators to differentiate between reverberations when it gets further away from the object casting the artifacts, leading to more difference in annotations. Semisupervised approaches can reduce human labeling time, and
weak supervision can estimate how much of the pixel is corrupted by the reverberation artifact.

In this paper, we propose a novel weakly- and semisupervised probabilistic needle and reverberation artifact segmentation algorithm and show that our algorithm is useful in improving the performance of downstream tasks. The algorithm can be broken into three parts. The first part is a probabilistic segmentation network trained with hard binary and not-so-accurate labels. The probabilistic segmentation is used to cope with the labeling errors and ambiguity in the input images. The second part is a transform function that turns the output of the first network into soft-labeled images where the value of each pixel represents how much of the pixel value is affected by the artifact. The last part is another network that is trained with the newly generated labels as desired output. The main contributions of this paper are (1) We developed a weakly- and semisupervised segmentation framework to segment the needle and reverberation artifacts which could cope well with the insufficiency in labeled data. (2) Our probabilistic approach is able to deal with ambiguity in the artifact boundaries and variations in human labels. (3) Our approach estimates how much of the pixel values are corrupted by reverberation artifacts.

2. Related Work

There already exists a long history of research algorithms to identify and remove reverberation artifacts from medical images. [6] produce a near optimal estimate of the reflectivity value due to reverberation by soft thresholding the 2D discrete wavelet transform. [7] improve the method proposed by [6] by utilizing radiofrequency (RF) data and soft thresholds on the wavelet coefficients to estimate the reflectivity values caused by the artifact. [8] model the reverberation mathematically, identifying and removing the artifact in RF data. Temporal information is used by [9] when they develop the 3D filter to reduce reverberation artifacts. These methods either made use of the RF data or temporal information, i.e., video sequence, but most of the times, we will
not have access to RF data or a large number of ultrasound videos.

Many deep learning segmentation frameworks for medical images can be good at segmenting reverberation artifacts too, but to the best of our knowledge, there is no learning-based segmentation framework specifically for the purpose of reverberation artifacts in ultrasound images. U-Net [10] use an encoder-decoder framework with skip connections between the encoder and the decoder, enabling the network to tend to more fine-grained details. [11] put Long Short-Term Memory layers (LSTM) and U-Net together, enabling the proposed USVS-Net to excel at identifying ambiguous boundaries and be robust against speckle noises. Attention U-Net is proposed by [12] to suppress irrelevant regions and has the network focus more on the target structure with different shapes and sizes. [13] recalibrate the channels by putting Squeeze and Excitation (SE) blocks [14] into a normal U-Net to increase the values of meaningful features and suppress the values of the insignificant ones. Together with dual attentive fusion blocks, adversarial learning is used in [15], allowing semisupervision of the segmentation, instead of full supervision, so that it is able to ameliorate the problem of not having enough labeled data. These methods are robust in a fully supervised situation and are good at segmenting structures with hard labels, but they are not able to generate segmentation masks that have a meaningful probabilistic output, i.e. they are unable to model the ambiguity in the data.

[16] explicitly model the aleatoric and epistemic uncertainty in deep learning and increased the performance on noisy data sets with dropout layers and an additional variance term in the output. [17] adopt Monte Carlo dropout to explore the uncertainty in the network for medical image segmentation. The work further explains how to capture different types of uncertainty in medical images. [18] combined a conditional variational autoencoder with a U-Net. The model does not just provide a single segmentation for a given input; instead, it predicts multiple plausible segmentations drawn from the distribution learned from the training data. However, this framework only works well on images with a single object or with global variations. As a result, a hierarchical probabilistic U-Net (HPU-Net) [19] is proposed to solve such problems. Instead of global latent variables, coarse-to-fine hierarchical latent variable maps are used in this work. Besides U-Net like blocks in the decoder side, some blocks are also sampled from the prior which is trained to have similar distribution with the posterior. [20] model the conditional probability distribution of the segmentation given an input image with a network named PHINet. Like the HPU-Net, PHINet also uses a posterior network and a prior network. However, unlike the HPU-Net, the likelihood network in this work completely samples from the prior network.

In terms of weakly-supervised learning methods, [21, 22] investigated the weakly supervised object detection problem by leveraging prior knowledge and context information. [23] discussed the techniques to train a deep saliency network without annotations with synthesized supervision. [24] phrased the training objective as a biconvex optimization for linear models which could be interpreted as a novel loss function with arbitrary linear constraints on the structured output space of pixel labels. [25] improved the performance of weakly-supervised semantic segmentation on natural images by a self-supervised difference detection module that estimates noise from the results. In the medical domain, [26] trained a convolutional neural network (CNN) with discrete constraints and regularization priors to approach this problem. [27] proposed a weakly-supervised framework to segment the shadows in ultrasound images. Essentially, the work uses one network for the initial segmentation, feeds the output of this network into a transform function, and then gets a new set of labels to train another network. The transform function in this paper transforms the hard labels to soft labels and, at the same time, reduces the effect of labeling error.

3. Method

3.1. Overview. Due to the difficulty in defining the actual boundaries of needles, our method is designed to deal with over-labeling from annotators. Over-labeling is a labeling scheme that includes all the pixels of interest in the label, but it would falsely include background pixels in the label as well. The goal of over-labeling is to have fewer false negatives, but the tradeoff is that it also includes more false positives. Over-labeling is much faster than the traditional accurate labeling while including more true positives. Inspired by [27], we proposed a multistep segmentation framework, shown in Figure 2. The entire framework can be divided into three parts: (1) a probabilistic segmentation trained on over-labeled hard labels, (2) a transformation function which takes in the output of the first network and transforms the probabilistic outputs into soft labels to remove the false positives and quantify the artifacts, and (3) another probabilistic segmentation network trained on the newly generated soft labels. Our algorithms differ from [27] in four main parts: (1) our transform function is designed based on the appearances and physical properties of reverberation artifacts instead of shadows, (2) we not only improve the segmentation masks in the transform function but also quantify the artifacts, (3) we propose to use probabilistic segmentation to compensate for the over-labeling by human experts instead of deterministic segmentation, and (4) we propose a new probabilistic segmentation network that learns variance from a known variance posterior. Even though the first part of the algorithm is fully supervised, accurate labeling is not needed, since the probabilistic segmentation method models the ambiguity in the labels and the original inputs. After getting the first network trained, the unlabeled ultrasound images can be passed through the network, creating hard segmentations on the unlabeled images. The amount of training data for the second network can also be increased with this approach. The output of the first network is then modified by a transform function, which converts the hard-labeled segmentation into soft labels. The transform function is designed specifically for the appearance and physical properties of reverberation artifacts in ultrasound images. The second network is
After that, we use the soft labels to train the soft label segmentation network. To train the hard label segmentation network, we transform the output of this network to soft labels by the transform function. To design the network, we use probabilistic networks and a novel transform function in our work. The pipeline can be divided into three parts: (1) probabilistic segmentation on hard labels, (2) transform function, and (3) probabilistic segmentation on soft labels. We use the hard labels to train the hard label segmentation network, then transform the output of this network to soft labels by the transform function. After that, we use the soft labels to train the soft label segmentation network.

Figure 2: The pipeline of the algorithm is inspired by [27]. However, due to the difference between reverberation artifacts and shadow artifacts, we use probabilistic networks and a novel transform function in our work. The pipeline can be divided into three parts: (1) probabilistic segmentation on hard labels, (2) transform function, and (3) probabilistic segmentation on soft labels. We use the hard labels to train the hard label segmentation network, then transform the output of this network to soft labels by the transform function. After that, we use the soft labels to train the soft label segmentation network.

designed to further eliminate the effects of uncertainty in human labels and ambiguity of the images, creating a soft-labeled segmentation which represents how much of each pixel value is corrupted by reverberation artifacts.

3.2. Probabilistic Segmentation on Hard Labels. The artifacts might differ by shape, intensity distribution, and unclear boundaries, so human labels could be different across annotators. Even when the same annotator labels the same image multiple times, the results can still differ. As a result, we conjecture that careful elimination of the ambiguity in the labeling can assist with subsequent analysis. Ideally, the segmentation algorithm needs to generate nearly identical results for the same image despite using data labeled by different annotators. The work in [19] models the ambiguity in the images and the labels, which is what we look for in this stage of work. Therefore, we utilize their network for our first step: segmentation on hard labels. In our network, even though we used the same general structure as [19], we used more local blocks than global blocks, so that we could model the ambiguity of the edges better.

We train the network with hand-labeled images and cross entropy as loss function, but after it is trained, we do not necessarily need to use the labeled training images for the following steps. Sampling from a learned distribution, the network will generate two mean maps and two standard deviation maps for the artifacts and needles segmentation. We denote these output mean maps for artifacts and needles as \( \mu_{\text{artifact,hard}} \) and \( \mu_{\text{needle,hard}} \), respectively. We denote the output standard deviation maps for artifacts and needles as \( \sigma_{\text{artifact,hard}} \) and \( \sigma_{\text{needle,hard}} \), respectively.

3.3. Transform Function. Even though the output of the first network inherently models the ambiguity in the input images and the human labels, it does not account for the fact that weaker reverberation artifacts do not affect the quality of the image as much as the strong ones do. Therefore, we transform the segmentation mask into a soft labeling mask, which models how much the artifact affects the pixel value. There are three components to our transform function: (1) reduction of false positives, (2) introduction of an exponential decay depending on how far the artifact is relative to the needle causing the artifact, and (3) lowering the artifact-segmentation soft labels for pixels in between adjacent reverberations.

The first part of the transform function takes in predicted hard needle mask \( \mu_{\text{needle,hard}} \) and artifact mask \( \mu_{\text{artifact,hard}} \). We begin by clustering the artifacts based on distance. We then determine if there is a needle closely above a cluster of artifacts: if there is, then keep the artifacts, and set the needle as the cause for this cluster of artifacts. This is to deal with the situation where there are multiple needles in the image. If not, then remove the cluster. The pseudocode for the algorithm is shown in Algorithm 3.3. The horizontal threshold \( ht \) will be small because needle artifacts are typically (near) continuous horizontal lines, whereas the vertical threshold \( vt \) will be larger to encompass the vertical spacing between artifact lines, which is based on the needle’s reverberating cross-section [2]. The threshold \( t \) indicates the largest possible distance between the segmented artifacts and the corresponding needles for the artifacts to be considered true positives. In this paper, for 256 \( \times \) 256 images, we set \( ht = 7 \), \( vt = 11 \), and \( t = 10 \). The hyperparameters should not affect the result much as long as they are in a reasonable range. The hyperparameter selection process is further discussed in Section 4.

The second part of the the transform function’s purpose is to create the exponential decay in the segmentation and, at the same time, compensate for the pixels that do not comply with the decay. As sound waves travel through a medium, the intensity falls off as the distance increases. Normally, the intensity decays exponentially [28], and due to the fact that reverberation artifacts are caused by sound waves reverberating inside the tissue [1], the intensity of the reverberation artifacts should be no different. The first bounce of the sound wave at the needle represents the needle, followed
Data: \( \hat{\mu}_{\text{needle hard}} \), \( \hat{\mu}_{\text{artifact hard}} \)

Result: \( y_1 \); Artifact mask with false positives removed

\[ size = \text{input image size} \]
\[ B = \text{zeros}(size), y_1 = \text{zeros}(size); k = 1; \]

For (row \( i \), column \( j \)) where \( \hat{\mu}_{\text{artifact hard}}[i,j] > 0 \) do

If \( B[i,j] == 0 \) then

Create stack \( s \);\n
\[ s.\text{push}([i,j]); B[i,j] = k; \]

While \( s \) is not empty do

\[ [x,y] = s.\text{pop}; \]

For \( [ii,jj] \) where \( (ii-x/\mu t)^2 + (jj-y/\mu t)^2 < 1 \) do

If \( B[ii,jj] == 0 \) then

\[ s.\text{push}([ii,jj]); B[ii,jj] = k; \]

\[ k++; \]

Set \( 1 = \text{where}(\hat{\mu}_{\text{needle hard}} > 0.5); \]

For \( (kk = 1; kk < k; kk++) \) do

Set \( 2 = \text{where}(B == kk); \]

If \( \exists(i_0,j_0) \in \text{set}1, \text{and} \forall[i_2,j_2] \in \text{set}1, \text{s.t.} i_0 \text{below} i_2 \text{and distance}([i_0,j_0], [i_1,j_1]) < t \) then

\[ y_1 += (B == kk) \ast \hat{\mu}_{\text{artifact hard}} \]

Algorithm 1: False Positive Removal, where size is the image size, \( B \) denotes the cluster each pixel belongs to (if a pixel does not belong to any cluster then its value in \( B \) would be 0), and \( y_1 \) is the output artifact mask with false positives removed. The outer loop with \( i \) and \( j \) iterates trough all the pixels that have a value larger than zero in the mask \( \hat{\mu}_{\text{artifact hard}} \), if the pixel \( (i,j) \) does not belong to any cluster (i.e., \( B[i,j] == 0 \)), then we push the pixel into a newly created stack \( s \) and set the cluster of the pixel to \( k \). Inside the loop, while the stack \( s \) is not empty, we pop out a pixel \((x,y)\) from the stack and search within an ellipse around it (i.e., \((ii-x/\mu t)^2 + (jj-y/\mu t)^2 < 1\), where \((ii,jj)\) is the pixel within the ellipse). If the \((ii,jj)\) does not belong to any cluster, then we also push this pixel onto stack \( s \) and set the cluster of the pixel to \( k \). After the stack \( s \) is empty, we increase \( k \) by 1, meaning that we move onto the next cluster for the next iteration of the outer loop. After all the pixels have been clustered, we continue to examine if a cluster \( kk \) is below and close enough to the needle. If it is, we include that cluster of artifacts in our output \( y_1 \).

by resonating bounces resulting in the artifact, so the intensity of the artifact should be on an exponential decay of the intensity of the needle. The intensity of the artifact falls until the last of the identified artifact pixels towards the bottom. We model this exponential decay as follows:

\[
y_{2,1}(i,j) = y_1(i,j) e^{-\frac{d(i,j)}{\alpha}}.
\] (1)

where \( \alpha \) is a hyper parameter, depending on the ultrasound imaging settings. The higher the frequency, the larger \( \alpha \) should be, as sound waves would then encounter more resistance in-depth. For the experiments in this paper, \( \alpha = 0.8 \).

\( h(i,j) \) represents the distance of pixel \((i,j)\) to the needle that is causing the artifact. \( d(i,j) \) denotes the distance between the deepest pixel (the furthest pixel away from the corresponding needle) in the cluster of artifacts containing \((i,j)\) and the nearest pixel in the corresponding needle.

However, other objects and tissues in the image would also have minor effects on the pixel values in the reverberation artifacts, e.g., other boundaries which overlap with the reverberation, shadows caused by vessels interacting with the reverberations, etc. The exponential-decay artifact model does not account for these other components of the pixel values. To deal with the problem, we also create an alternate measurement \( y_{2,2} \) based on the pixel values in the input images. Denote the input image as \( I \). For normalization, we first find the maximum pixel value \( m_1 \) in the needle-region of \( I \). The normalized pixel values in \( I \) can then be used as weights on the artifact soft-label mask as follows:

\[
m_1 = \max_{\forall(i,j), \hat{\mu}_{\text{needle hard}}(i,j) > 0.5} I(i,j),
\]

\[
y_{2,2}(i,j) = \frac{I(i,j)}{m_1} y_1(i,j).
\] (2)

In cases where artifact pixels are unusually bright, e.g., \( y_{2,2} \) is large due to overlapping with actual object boundaries. Preserving such property is desired because it represents the actual anatomy or a different artifact (such as a diagnostic B-line in a lung). We therefore combine \( y_{2,1} \) and \( y_{2,2} \) by taking the maximum:

\[
y_2(i,j) = \max \left( y_{2,1}(i,j), y_{2,2}(i,j) \right).
\] (3)

At this point, we have removed the false positives and also created an exponential decay in the artifact soft labels while preserving the effects of underlying anatomic boundaries on artifacts. However, we still need to assign lower values to the soft labels in between each reverberation, since the hard human labels tend to (incorrectly) inpaint such regions. The pixels between reverberations tend to have lower values than the artifact pixels in the original input images. In the third part of our transform function we determine whether a certain pixel belongs to the
nonartifact region in between reverberation lines. We need to compare the value of that pixel in the original ultrasound image against the maximum value in a local patch. We also apply a sigmoid-like function to make sure that artifact soft labels are limited with respect to the highest intensity value within that region. Intuitively, the sigmoid-like function ensures that the values in \( y_2 \) which are close to the local maximum (which should be real artifact pixels) would be mostly preserved and the values close to zero (which should not be artifact pixels) would be mostly removed, while the values in between follow a smooth decay in the output artifact mask. The equations are listed as follows:

\[
m_2(i,j) = \max_{(i_1,j_1) \in \Omega_2(vw,hw)} I(i+i_1, j+j_1),
\]

\[
H_{\text{artifact, soft}}(i,j) = \frac{y_2(i,j)}{1 + e^{-\beta(i,j)/m_2(vw,hw)}}.
\]

where \( \beta \) is a hyper-parameter that controls how fast the fall-off is. If the noise level is high, a larger \( \beta \) should be used. Also, \( \Omega_2(vw,hw) \) is a rectangular region where \((0,0)\) is the center point, and \(2vw\) and \(2hw\) is the height and width. \( vw \) and \( hw \) stand for vertical and horizontal width, respectively. \( vw \) should be large enough to include at least one line of true reverberation artifact in the patch. In this paper, we set \( \beta = 8, vw = 2, \) and \( hw = 1. \) Again, these parameters are not sensitive, as long as \( vw \) and \( hw \) stay relatively small and more details are discussed in Section 4.

As for the standard deviation map of artifacts, we want to rescale its values in the same manner as we rescale the mean map of artifacts. Therefore, the transform function for the standard deviation map can be simplified to

\[
\sigma_{\text{artifact, soft}}(i,j) = \sigma_{\text{artifact, hard}}(i,j) \frac{H_{\text{artifact, soft}}(i,j)}{H_{\text{artifact, hard}}(i,j)} + \epsilon.
\]

where \( \epsilon \ll 1 \) avoids division by zero.

Since needles are a lot more visible and less ambiguous than reverberation artifacts and the needle boundaries are better defined than the artifact boundaries, the probabilistic output of the first network would be enough for our purposes. Therefore, the needle labels are not processed, thus will not change in this part, as a result, we have \( H_{\text{needle, soft}}(i,j) = \mu_{\text{needle, hard}}(i,j), \) and \( \sigma_{\text{needle, soft}}(i,j) = \sigma_{\text{needle, hard}}(i,j). \)

Among the hyperparameters mentioned in the subsection, the values of \( ht, vt, t, vw, \) and \( hw \) are selected based on the observation of our particular dataset. In other words, the values for these hyperparameters are dependent on the appearance of the needles and artifacts, as well as the scale (i.e., physical distance between pixels). These hyperparameters are more related to the imaging subject and the scale of the image rather than the imaging setting. On the contrary, \( \alpha \) and \( \beta \) are more dependent on the imaging setting, as \( \alpha \) indicates the sound waves’ ability to penetrate and the selection of \( \beta \) is affected by the noise level in the image.

### 3.4. Probabilistic Segmentation on Soft Labels

Our segmentation model here is built upon the HPU-Net [19], and we follow their notation below. We want to model the unknown ambiguity in the images and the labels, and at the same time, we also want to take the known variance in the labels into account. Our training objective is similar to that of the previous work: maximizing the evidence lower bound on the likelihood \( p(M|X) \), except that we are modeling a variational posterior \( Q(\Omega,|X,M,V) \) instead of \( Q(|X,M) \), where \( X \) is the input image, \( M \) is the known mean of the segmentation label, and \( V \) is the variance of the segmentation label. Denote \( z_{cj} = \{ z_{1i}, z_{2i}, \ldots, z_{ni} \} \). We calculate the posterior \( Q \) from two separate networks, where one network accounts for the mean \( \mu_{i}^{\text{post}}(z_{cj}, X, M) \) and the other network models the variance \( \sigma_{i}^{\text{post}}(z_{cj}, X, V) \). The latent features in the prior blocks should follow a normal distribution generated by the posterior blocks \( N(\mu_{i}^{\text{post}}(z_{cj}, X, M), \sigma_{i}^{\text{post}}(z_{cj}, X, V)) \). During training, we directly sample from the posterior \( Q \) and train the normal distribution generated by the prior \( N(\mu_{i}^{\text{prior}}(z_{cj}, X), \sigma_{i}^{\text{prior}}(z_{cj}, X)) \) to be close to the one from the posterior. When sampling, we did exactly the same as the previous work did: sampling the latent features from the normal distribution modeled by the prior. The training and sampling process is illustrated in Figure 3.

In this particular application, we utilize a mean-squared-error-based custom loss function as a way to deal with the continuous values and unique meaning of soft labels. To deal with overfitting to the background, we only care about the pixels that have values over a certain threshold \( \gamma \), which we set to 0.05. Lower weights are assigned to pixels where absolute error is within the known standard deviation, since we are less sure about the value in the label where standard deviation is larger. Therefore, the loss function can be expressed as the following:

\[
\text{loss} = \sum_{(i,j) \in \Omega_2} w(i,j) \left( \bar{\mu}_{\text{soft}}(i,j) - \mu_{\text{soft}}(i,j) \right)^2,
\]

where

\[
\Omega_2 = \{ (x,y) | \bar{\mu}_{\text{soft}}(x,y) > \gamma \vee \mu_{\text{soft}}(x,y) > \gamma \},
\]

\[
w(i,j) = \begin{cases} 
  k & \text{abs}(i,j) < \sigma_{\text{soft}}(i,j) \\
  1 & \text{abs}(i,j) \geq \sigma_{\text{soft}}(i,j)
\end{cases},
\]

\[
\text{abs}(i,j) = |\bar{\mu}_{\text{soft}}(i,j) - \mu_{\text{soft}}(i,j)|,
\]

\[
k < 1.
\]

### 4. Experiments

The ultrasound imaging is performed with a UF-760AG Fukuda Densihi diagnostic ultrasound imaging machine, on both a chicken breast and an anthropomorphic phantom containing simulated vessels in a tissue medium produced by Advanced Medical Technologies. We used a linear transducer with 51 mm scanning width. It was set to 12 MHz with
an imaging depth of 4 cm and a gain value of 21 dB. The dimension of each image is $512 \times 532$ pixels, and the pixel pitch is 0.10 mm horizontally and 0.075 mm vertically. The images were then resized to $256 \times 256$. All the networks are trained on a single Nvidia Titan RTX for 15 epochs with horizontal flipping, Gaussian blurring ($\sigma \sim 0.5, 1$, where $\sigma \sim 0$ is taking the original image) and Gamma transform ($y = 0.9, 1, 1.1$) as data augmentation, adam [29] as optimizer with a learning rate of $1 \times 10^{-3}$. We note that one training epoch consists of all possible 18 (2 for flipping $\times 3$ for Gaussian blurring $\times 3$ for Gamma transform) combinations of data augmentation. Then, we picked the best model based on the validation results. There are 401, 100, and 40 images in the training, validation, and test set, respectively. In other words, 7218 training images were used to train each epoch due to data augmentation. Our dataset is split in this ratio due to the sophisticated labeling of the test set. The labeling in this work utilized the labeling tool Labelme [30]. The training of the first and second network in our framework both took about 90 minutes. The transform function was performed on Intel Core i5-8279U which takes about 80 minutes to process 401 images.

The training set is over-labeled (shown in the second images of Figure 1(a)) to show that our algorithm is good at dealing with inaccurate labels as well as with handling false positives in the training data. Since there is not a clear definition of the boundaries and quantification of reverberation artifacts, traditional evaluation metrics, e.g., Dice coefficient and Intersection over Union (IoU) would not work here. Therefore, we carefully label the test set in the following way for better evaluation. An example of our test set labeling is shown in the left image of Figure 1(b), where the yellow and pink regions indicate the possible areas containing needles and artifacts, respectively, green and white regions represent the first and second reverberations that we are confident in, the bright blue circles represent the patch between reverberation with no artifacts, the dark blue represents the needle patches that we are confident in, and the red patches are the regions with identifiable fuzzy-artifacts. We define the artifact and needle pixels with values greater than 0.05 and 0.5 as positive, respectively. The labeling is designed to show that in our result, (1) the pixels that definitely belong to needles and to the significant artifacts are segmented, (2) the pixels that are between the reverberations and not part of the artifacts are not segmented, (3) the fuzzy artifacts are segmented with lower values, and (4) very few false positives are included.

Our notations are summarized in Figure 1(c). We denote the ratio between positive artifact and needle counts in the labeled first, second reverberations and needle and the total pixels counts in those regions as FAR (first artifact rate), SAR (second artifact rate), and NR (needle rate), respectively, and the average value for positives in those artifact patches as FAA (first artifact average) and SAA (second artifact average). We also denote the false positive rate of artifacts and needles as AFPR (artifact false positive rate) and NFPR (needle false positive rate), and the average value of the false positives as AFPA (artifact false positive average) and NFPA (needle false positive average). We also calculate the average of the nonartifact region between reverberations as NAA (nonartifact average). We further define the average positive artifact value in the identifiable fuzzy artifact patches as IFAA (identifiable fuzzy artifact average).

The following paragraph describes the intuition behind the proposed measurements. By design, FAR, SAR, and NR measure whether the segmentation algorithm is able to segment the pixels that absolutely belong to the first reverberation, the second reverberation, or the needles, respectively, so the closer they are to 1, the better the segmentation algorithm performs. Next, we examine the FAA, SAA, NAA, and IFAA measurements together. Our intuition is that the second reverberation (white region in Figure 1(b)) is deeper in position and darker in brightness than the first one (green region in Figure 1(b)), so the second reverberation corrupts the “real” values of the pixels less. The reverberations below the second (identifiable fuzzy artifacts, red region in Figure 1(b)) are usually more ambiguous and darker, so they should corrupt the pixel values less. To add to that, the nonartifact patch between reverberations (bright blue region in
Figure 1(b)) should have lower values than both the reverberations above and underneath it, as those pixels should not be a part of the reverberation artifacts. As a result, FAA should be larger than SAA, SAA should be larger than IFAA, and IFAA should be significantly larger than NAA, because the output values describe how much the artifacts corrupt the pixels. In other words, a good quantification model should produce $FAA > SAA > IFAA > NAA$. In terms of AFPR and NFPR, anything outside the pink and yellow regions in Figure 1(b) should be false positives, as they are clearly not needles or artifacts. Therefore, the AFPR and NFPR values should be zero in an ideal segmentation. Even though there might be false positives in the segmentation method, different false positives might have different consequences on the downstream tasks. False positives with really small quantification values would not be as bad as the ones with large output values. Therefore, we also investigate AFPA and NFPA, whose values represent the average value of the false positives in the output. The lower these values are, the less impact the false positives would have on the subsequent algorithms.

Before any quantitative experiments, we want to show some of the intermediate results from the transform function in order to justify the operation. As shown in Figure 4(a), it can be seen that the first part of the transform function removes the false positives in the output of the first network $\hat{\mu}_{\text{artifact,hard}}$, and the second part of the network creates the exponential decay while picking up the bright artifacts in regions that are further down, while the third part differentiates the reverberation from the background.

Furthermore, we also want to show that the hyperparameters are not very sensitive with a few examples shown in Figure 4(b), where we show the intermediate results of the transform function. We multiply all the hyperparameters in the transform function by a coefficient of 0.67 and 1.5 and compare the results against the original values we set. This simple experiment should be able to show that the hyperparameters are not sensitive to changes since these hyperparameters are independent of each other. It can be observed that there is not much difference in the final output of the transform function. Thus, the final results are not very sensitive to the choice of hyperparameters.

In our first quantitative experiment, we compared our results against those from expert labels, U-Net [10], USVS-Net [11], and HPU-Net [19] to show that our proposed algorithm is superior to these current segmentation networks and expert labels for needle-and-reverberation-artifact segmentation. The experts tend to label the entire region that might contain artifacts, but they do not differentiate between the reverberations. The quantitative results can be found in Figure 5(a). As can be seen in the table, our approach performs similarly to HPU-Net at binary-thresholded segmentation of the significant artifacts and needles, and our approach performs better than the other two methods. Our goal, however, is more than just binary segmentation. Since the value in the output represents how much the artifact affects the underlying “real” pixel values, and there should be a decay in the intensity of the artifacts, the average value of first artifacts should be larger than that of second artifacts. Together, they also should not be too close to 1 as they normally do not completely obscure the real tissue there. Our method is the only algorithm that has achieved such results. Our algorithm also limits the false positive rate to around 0.02, whereas the best results achieved by other algorithms are 0.084 and 0.04, for artifacts and needles, respectively. Our algorithm also does a good job in differentiating between reverberations, since the average value in nonartifact patches between reverberations is 0.042, which is the lowest by far, compared to 0.506, 0.431, and 0.136 from U-Net,
USVS-Net, and HPU-Net, respectively. The IFFA value further indicates that our network learns the decay in the artifact intensity. Sample results of needle segmentation and artifact segmentation can be found in right two image of Figure 1(b). Qualitative results of artifact segmentation are shown in Figure 5(c). Our method clearly does the best in differentiating between the actual artifacts and the patches around them, while simultaneously assigning lower values to less significant artifacts.

Our second quantitative experiment is done to show that our proposed second network in the pipeline works better than other networks. In this experiment, we test out U-Net, HPU-Net, and USVS-Net as the second network. We found that USVS-Net overfits to the background too much, so it is not included in the quantitative comparison. The quantitative evaluation can be found in Figure 5(b). It shows that our method is as good as HPU-Net in detecting the strong artifacts and needles, since the FAR, SAR, and NR values are similar. It also illustrates that our method works better in differentiating between the artifacts and the patches between reverberations, as well as giving higher values to the first reverberation than to the second one. Lastly, our
methods also produce fewer false positives and has lower values in the false-positive patches.

Our next experiment is to show that we can further improve our results with additional unlabeled data, and that if the training data of the first network is unavailable, using only nonlabeled data also would not hurt the performance much. As for the unlabeled data used in this study, there are 480 unlabeled images from the same phantom and chicken breast as the labeled training data, but the unlabeled data are from different trials than the labeled ones. We train

|        | Labeled only | Both   | Unlabeled only |
|--------|--------------|--------|----------------|
| FAR    | 0.986        | 0.988  | 0.971          |
| SAR    | 0.994        | 0.999  | 0.996          |
| NR     | 0.981        | 0.966  | 0.941          |
| FAA    | 0.728        | 0.723  | 0.698          |
| SAA    | 0.698        | 0.695  | 0.691          |
| AFPR   | 0.022        | 0.029  | 0.043          |
| NFPR   | 0.018        | 0.005  | 0.024          |
| AFPA   | 0.151        | 0.143  | 0.125          |
| NFPA   | 0.627        | 0.624  | 0.643          |
| NAA    | 0.042        | 0.033  | 0.040          |
| IFAA   | 0.471        | 0.440  | 0.497          |

|        | Hand labels  | Uncertainty-pruned labels |
|--------|--------------|---------------------------|
| FAR    | 0.986        | 0.958                      |
| SAR    | 0.994        | 0.997                      |
| NR     | 0.981        | 0.955                      |
| FAA    | 0.728        | 0.630                      |
| SAA    | 0.698        | 0.668                      |
| AFPR   | 0.022        | 0.063                      |
| NFPR   | 0.018        | 0.034                      |
| AFPA   | 0.151        | 0.137                      |
| NFPA   | 0.627        | 0.621                      |
| NAA    | 0.042        | 0.044                      |
| IFAA   | 0.471        | 0.400                      |

Figure 6: (a) A table showing the benefit from adding labeled or unlabeled data. We note that the first network is trained with the same training data in this experiment, and the unlabeled data are from the same subject as the training data but from different trials. (b) A table showing the robustness of our method across various binary labels. (c) Example of aleatoric uncertainty map and uncertainty-pruned label. Left to right: input image, aleatoric uncertainty map, hand label, and uncertainty-pruned label. (d) Example of failure cases. Left to right: input images, hand labels (where needles are gray and artifacts are white), needle outputs, and artifact outputs. It can be seen that when artifacts are less visible, the model fails to detect both needles and artifacts. Besides, small values are present in the artifact output where there are patterns that look similar to reverberation artifacts.
the first network with the same training data in this experiment, and then we define 3 different scenarios (1) we only have the labeled data to train the second network, (2) we have both labeled and unlabeled data for the second network, and (3) for some reason, the labeled training data is not available any longer so we only have the unlabeled data to train the second network. The comparison can be found in Figure 6(a), where it shows that using both labeled and unlabeled data slightly decreases the false positives of needles and also lowers the values between reverberations. On the other hand, using only unlabeled data produces only slightly more false positives.

To provide further validation of our algorithm on different sets of plausible hand labels, we apply aleatoric uncertainty maps obtained from a Monte Carlo dropout-based Bayesian 3-D convolutional neural network similar to [31]. In doing so, we are able to simulate the wave-like intensity decrease of needle artifacts in a realistic manner. For modeling the aleatoric uncertainty in the artifact segmentation region, we apply the model

\[ \frac{1}{T} \sum_{t=1}^{T} \text{diag}(\tilde{p}_t) - \tilde{p}_t^{\otimes 2}, \]  

using the notation from [31], where \( \tilde{p}_t = p(\tilde{w}_t) \) represents softmax outputs evaluated with a set of random weights \( w \), \( t \) represents each run of the Monte Carlo sampling, and \( \otimes \) represents the Kronecker matrix product. In an attempt to obtain more articulate hand labels, we apply the following steps on top of our hand labels: (1) use the labels to mask only the relevant regions in the aleatoric uncertainty map, (2) take the inverse of the uncertainty map, and (3) determine a local threshold \( \delta_{\text{thresh}} \) for smaller patches within each labeled region and use that for deciding whether to keep or remove the labels. A sample of the aleatoric uncertainty map and uncertainty-pruned label are shown in Figure 6(c). We compare the results from networks trained on the original hand labels against the ones trained on the refined labels. The results are shown in Figure 6(b). Our algorithm still does a decent job on the uncertainty-pruned labels, even though the performance is slightly worse than on original hand labels. The intuition is that although the uncertainty-pruning removes some of the labels in the spaces between reverberations. Our algorithm is designed to differentiate the reverberations from the spaces between reverberations to compensate for the over-labeling, so it does not make much difference in that regards. However, the pruning step removes some of the true positives as well, which could not be learned in the networks, slighting hurting the performance. The algorithm still does a decent job on a different set of labels, while performing slightly better on the over-labeled annotations.

However, our method sometimes fail when there are no obvious artifacts. Shown in Figure 6(d), on the right side of the two images, the reverberation artifacts are either not present or not obvious and both the segmentation of the needles and artifacts are not able to detect anything in the region. We suspect that the network learns to segment the needles by looking at the artifacts and not the other way around, resulting in needles undetected when there are no artifacts or when artifacts are less visible. This is probably due to the fact that most of the reverberations artifacts in our training dataset are very visible, so the network would
have bias towards this scenario. Besides, when there are patterns in the image that look similar to those of reverberation artifacts, the network tends to generate small values for those pixels on the artifact mask.

5. Applications

In this section, we show that our reverberation artifact segmentation and quantification method is useful in several downstream tasks, such as multiview image compounding and vessel segmentation. All of the imaging parameters are set exactly as they are in Section 4.

5.1. Multiview Image Compounding

The goal of multiview image compounding is to take the information from images taken at different viewpoints and reconstruct the true underlying structure. This task is important in ultrasound, for which images are path-dependent leading to certain structures being seen in an image taken from one viewpoint but not seen in a different image containing the same structure from a different viewpoint. However, the same object can cast reverberation artifacts in different directions in images from different viewpoints, making multiview compounding a challenging task. In compounding, we want to preserve the real objects and structures in the compounded image but remove the artifacts at the same time.

We apply our reverberation artifact segmentation and quantification algorithm in an image compounding task, to remove the reverberation artifacts in the compounded image. We propose to use a simple compounding algorithm that considers our reverberation artifact segmentation and quantification. For example, we can take two viewpoints—

\[
\text{for every pixel } (i, j), \text{ if } C_1(i, j) - C_2(i, j) > t_c, \text{ then we set } \hat{I}(i, j) = I_1(i, j); \text{ if } C_2(i, j) - C_1(i, j) > t_c, \text{ then we set } \hat{I}(i, j) = I_2(i, j); \text{ else, we set } \hat{I}(i, j) = \max(I_1(i, j), I_2(i, j)), \text{ where } t_c \text{ is a confidence threshold which we set as 0.1 in this paper. In other words, if the}
\]

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|c|}
\hline
 & Precision & Recall & IoU & Dice \\
\hline
U-Net & 0.823 & 0.801 & 0.684 & 0.812 \\
partial U-Net & 0.847 & 0.803 & 0.701 & 0.825 \\
\hline
\end{tabular}
\caption{Quantitative comparison between Partial U-Net and U-Net.}
\end{table}
confidence from one viewpoint is significantly higher, we take the pixel value from that viewpoint. Otherwise, we take the maximum image intensity across different viewpoints.

As an experiment, we take pairs of ultrasound images from orthogonal viewpoints with a needle inserted into the phantom. Since we are performing the experiment on a square phantom, it is easy to make sure the viewpoints are orthogonal with free-hand imaging. We compare our simple method against existing compounding algorithms, such as average (avg) [32], maximum (max) [33], and uncertainty-based fusion (UBF) [34]. The results are shown in Figure 7. We can observe that our simple compounding method preserves the vessels and needles as good as max [33] while nearly removing all the reverberation artifacts.

5.2. Vessel Segmentation. Ultrasound vessel segmentation is important for diagnostic purposes and for guiding needle insertion by both humans and robotic systems [35]. However, needle reverberation artifacts could occlude the structures or objects of interest [3]. Such artifacts could also affect the performance of vessel segmentation [5] and needle tracking algorithms [4].

We believe our reverberation artifact segmentation and quantification algorithm could help with vessel segmentation when reverberation artifacts produced by needles are present in the image. During convolution, segmentation networks should pay less attention to the artifact pixels, so we propose to use the soft segmentation results from our algorithm as the masks in partial convolution [36]. Although the partial convolution method is built for image inpainting, the idea to mask certain regions during convolution is useful for our purpose, since we do not want the network to treat the artifact pixels equally as other pixels.

In this work, we use U-Net [10] as the backbone for the vessel segmentation. We compare our artifact-segmentation- and quantification-based partial convolutional U-Net (partial U-Net) against a normal U-Net. For fair comparison, we have 5 layers in the encoder side, with 32, 64, 128, 256, and 512 filters for both the partial U-Net and the normal U-Net, and the decoder side is exactly symmetric to the encoder. We use Adam optimizer [29] with a learning rate of $1 \times 10^{-4}$ and binary cross entropy as loss function. There are 481 images in the training set and 97 images in both validation and test set. The only augmentation technique used during training is horizontal flipping.

Shown in the qualitative results in Figure 8(a), the results by normal U-Net create some false positives at the top left corner, where the needles are present. The needles together with the image boundaries or some other artifacts confuse the normal U-Net, since they look like vessel boundaries to some extent, while the partial U-Net does not generate false positive there. Quantitative results can be found in Figure 8(b). The precision for artifact-masked partial U-Net is notably higher than normal U-Net, indicating that our reverberation artifact quantification helps reduce the false positives in vessel segmentation. By other metrics, the partial U-Net also outperforms normal U-Net with the help of our soft reverberation artifact segmentation.

6. Discussion

Our proposed algorithm outperforms other existing segmentation algorithms in segmenting and quantifying reverberation artifacts. Our weakly- and semisupervised reverberation segmentation framework is able to take as input inaccurate human labels and predict accurate segmentation maps as well as quantitatively estimate how much the artifacts affect the pixel values. It works even better on over-labeled annotations. Our quantification of reverberation artifacts is useful in image compounding and for improving vessel segmentation results. Our probabilistic outputs could also potentially be used in ultrasound image uncertainty measurements, ultrasound image quality evaluation, and reverberation artifact removal. Lastly, this algorithm is also directly related to robot-controlled needle insertion, in which our work can be used to guide where to insert the needle, monitor the tissue under the needle, and determine the quality of the image.

Data Availability

The code and ultrasound data used to support the findings of this study are available at https://github.com/aL3x-O-o-Hung/reverberation-artifact-segmentation-and-quantification.

Disclosure

Please note that patents may be pending, and patent rights are not included in the CC by license.

Conflicts of Interest

Alex Ling Yu Hung is currently a student at the Department of Computer Science and the Department of Radiological Sciences, UCLA. Edward Chen is currently a student at Computer Science Department, Stanford University. The work was done when they were students at Carnegie Mellon University. They have no other affiliations or relationship with other organizations. John Galeotti has financial, consulting, management, and advisory relationship with Activ Surgical, which (currently) deals with surgical video rather than ultrasound and Elio AI, Inc., which is developing ultrasound AI and is licensing this technology. Galeotti serves on the advisory board for Activ Surgical, Inc., and he is a founder and director for Elio AI, Inc.

Acknowledgments

This present work was sponsored in part by the US Army Medical contracts W81XWH-19-C0083, W81XWH-19-C0101, and W81XWH-19-C-0020 and by the funding from the state of Pennsylvania C00072473. We would like to thank our collaborators at the University of Pittsburgh, Triton Microsystems, Inc., Accipiter Systems, Inc., Sonivate Medical, Inc., and URSUS Medical, LLC.
References

[1] M. Ziskin, D. Thickman, N. Goldenberg, M. Laypowker, and J. Becker, "The comet tail artifact," Journal of Ultrasound in Medicine, vol. 1, no. 1, pp. 1–7, 1982.

[2] R. M. Kirberger, "Image artifacts in diagnostic ultrasound—a review," Veterinary Radiology & Ultrasound, vol. 36, no. 4, pp. 297–306, 1995.

[3] J. Mohebali, V. I. Patel, J. M. Romero et al., "Acoustic shadowing impairs accurate characterization of stenosis in carotid ultrasound examinations," Journal of Vascular Surgery, vol. 62, no. 5, pp. 1236–1244, 2015.

[4] X. Xu, Y. Zhou, X. Cheng, E. Song, and G. Li, "Ultrasound intima-media segmentation using Hough transform and dual snake model," Computerized Medical Imaging and Graphics, vol. 36, no. 3, pp. 248–258, 2013.

[5] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in International Conference on Medical image computing and computer-assisted intervention, pp. 234–241, Springer, Cham, 2015.

[6] T. S. Mathai, V. Gorantla, and J. Galeotti, "Segmentation of vessels in ultra high frequency ultrasound sequences using contextual memory," in International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 173–181, Springer, Cham, 2019.

[7] O. Oktay, J. Schlemper, L. L. Folgoc et al., "Attention u-net: Learning where to look for the pancreas," 2018, http://arxiv.org/abs/1804.03999.

[8] A. G. Roy, N. Navab, and C. Wachinger, "Concurrent spatial and channel ‘squeeze & excitation’in fully convolutional networks," in International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 421–429, Springer, Cham, 2018.

[9] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 192.2–1931, Salt Lake City, Utah, 2018.

[10] L. Han, Y. Huang, H. Dou et al., "Semisupervised segmentation of lesion from breast ultrasound images with attentional generative adversarial network," Computer Methods and Programs in Biomedicine, vol. 189, article 105 275, 2020.
[32] J. W. Trobaugh, W. D. Richard, K. R. Smith, and R. D. Bucholz, "Frameless stereotactic ultrasonography: Method and applications," *Computerized Medical Imaging and Graphics*, vol. 18, no. 4, pp. 235–246, 1994.

[33] A. Lasso, T. Heffter, A. Rankin, C. Pinter, T. Ungi, and G. Fichtinger, "Plus: open-source toolkit for ultrasound-guided intervention systems," *IEEE transactions on biomedical engineering*, vol. 61, no. 10, pp. 2527–2537, 2014.

[34] C. S. Z. Berge, A. Kapoor, and N. Navab, "Orientation-driven ultrasound compounding using uncertainty information," in *International conference on information processing in computer-assisted interventions*, pp. 236–245, Springer, Cham, 2014.

[35] J. Guerrero, S. E. Salcudean, J. A. McEwen, B. A. Masri, and S. Nicolaou, "Real-Time vessel segmentation and tracking for ultrasound imaging applications," *IEEE Transactions on Medical Imaging*, vol. 26, no. 8, pp. 1079–1090, 2007.

[36] G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro, "Image inpainting for irregular holes using partial convolutions," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 85–100, Munich, Germany, 2018.