ROBUST BONE SURFACE AND ACOUSTIC SHADOW SEGMENTATION FROM ULTRASOUND FOR COMPUTER ASSISTED ORTHOPEDIC SURGERY

By

AHMED ZAKI ALSINAN

A dissertation submitted to the

School of Graduate Studies

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Electrical and Computer Engineering

Written under the direction of

Ilker Hacihaliloglu

And approved by

_____

_____

_____

_____

New Brunswick, New Jersey

January 2021

ABSTRACT OF THE DISSERTATION

Robust Bone Surface and Acoustic Shadow Segmentation from Ultrasound for Computer Assisted Orthopedic Surgery

by Ahmed Zaki Alsinan

dissertation Director: Prof. Ilker Hacihaliloglu

Orthopedic surgeries have been a prominent procedure in treating interminable pain and disabilities, due to musculoskeletal diseases, e.g. osteoarthritis, spinal conditions, osteoporosis, and low-energy fractures. Imaging has been an integral component of surgical and non-surgical orthopedic procedures such as total knee replacement (TKR), intramedullary nail locking for femoral shaft fractures, and pedicle screw insertion for spinal fusion surgery. Current practice during these procedures relies on intra-procedure 2D fluoroscopy as the main imaging modality for localization and visualization of bones, fractures, implants, and surgical tool positions. However, with such projection imaging, surgeons and clinicians typically face considerable difficulties in accurately localizing bone fragments in 3D space and assessing the adequacy and accuracy of the procedure. This problem has been overcome with 3D fluoroscopy units, however, they are twice as expensive and not widely available as standard 2D units. Additionally, fluoroscopy involves significant ionizing radiation exposure, which should be kept at minimal in order to avoid potential long-term complications. In order to overcome some of these limitations and provide a safe alternative, 2D/3D ultrasound (US) has emerged as a safe alternative while remaining relatively cheap and widely available. US image data, however, is typically characterized by high levels of speckle noise, reverberation, anisotropy and signal dropout which introduce sig-

nificant difficulties during interpretation of captured data. Limited field-of-view and being a user dependent imaging modality cause additional difficulties during data collection since a single-degree deviation angle by the operator can reduce the signal strength by 50%. In order to overcome these difficulties automatic bone segmentation and registration methods have been developed.

The goal of this research is to develop robust, accurate, real-time and automatic image segmentation and localization methods for bone structures in US guided interventional orthopedic procedures. A multimodal convolutional neural network(CNN)-based technique is developed for segmenting bone surfaces from in vivo US scans, in which fusion of feature maps and multimodal images are incorporated to abate sensitivity to variations that are caused by imaging artifacts and low intensity bone boundaries. A block-based CNN for segmentation of bone surfaces from in vivo US scans is also proposed. We utilize fusion of feature maps and employ multi-modal images to abate sensitivity to variations caused by imaging artifacts and low intensity bone boundaries. We also propose a conditional Generative Adversarial Network (cGAN)-based method for accurate real-time segmentation of bone shadow regions from in vivo US scans. Finally, a novel GAN architecture designed to perform accurate, robust and real-time segmentation of bone shadow images from in vivo US data, is proposed. We show how the segmented bone shadow regions can be used as an additional proxy to improve bone surface segmentation results of a multi-feature guided CNN architecture. Extensive validation studies were performed to address the engineering challenges found in real clinical situations. Validating the proposed methods with clinical studies will in turn help in the future design, development, and evaluation of a 3D US based CAOS system which could improve performance by providing better assessment and placement and results in reduction of operations time. This can decrease the cost and improve efficiency by replacing fluoroscopy at key points in the diagnosis and treatment.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my advisor Dr. Ilker Hacihaliloglu for his motivation, guidance and tremendous support throughout my research journey here at Rutgers. Not only his knowledge and expertise in the fields of medical image processing, orthopedics and ultrasound that made this work possible, but also his inspiring way of thinking about research problems. I would like to also express my sincere gratitude to my co-advisor Dr. Vishal M. Patel for his continued support and guidance throughout the years. I thank them both and my dissertation committee members: Dr. Zoran Gajic, Dr. Laleh Najafizadeh and Dr. Claudia Errico for their encouragement, motivation and constructive feedback. I would like to also thank Dr. Gajic for being truly an amazing director to our ECE department. His advice and guidance to me since day one made all the difference and I greatly appreciate it.

Many thanks go to my collaborators Charles Rule and Dr. Michael Vives and to all colleagues in Dr. Hacihaliloglu's CompAST lab and Dr. Patel's VIU lab. I am immensely grateful to Cosmas Mwikirize and David Le for their friendship, encouragement, and discussions. I have been fortunate to be able to attend Rutgers and Michigan State for many years and learn from remarkably inspiring professors. I would like to thank Lawrence R. Rabiner, Sophocles Orfanidis, George V. Moustakides, Kristin Dana, Hana Godrich, John J. McGarvey, Vivek Singh, Lalita Udpa, Satish Udpa, Ramakrishna Mukkamala, Hayder Radha, Donnie Reinhard, Gregory Wierzba and my mentors Michael S Morris and Tariq Khan.

Lastly, I would like to thank my parents Iftekhar Abu Alsaud and Zaki Alsinan for always believing in me. I am grateful to my wife, Zahraa, for her love and care to me and our daughter Sarah. I would like to also thank my parents-in-law Faizah Aljishi and Hilmi Alsinan for their support and encouragement.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

1.1 Thesis Motivation and Problem Statement

Orthopedic surgeries have been a prominent procedure in treating interminable pain and disabilities, due to musculoskeletal diseases, e.g. osteoarthritis, spinal conditions, osteoporosis, and low-energy fractures. In 1990, the World Health Organization reported 1.7 million hip fractures, for example, and projected the figure to increase to 6 million by 2050 [1]. Osteoporosis and related fracture treatments costs were estimated at $\$19.1$ billion in 2004. Moreover, spine related injuries for the years 2002-2004 were estimated at $\$193.9$ billion [1, 2]. Surgical interventions may include osteotomy, fracture fixation, or placement of an implant device [3]. The need for high precision in the mentioned procedures is evidently required in order to minimize the intra- and post-operative complications. Imaging has been an integral component of any orthopedic surgery. Intra-operative fluoroscopy and preoperative computed tomography are the most common imaging modalities in such surgeries. Despite their high-quality display of bone structures, fluoroscopy, and computed tomography (CT)-guided orthopedic surgery present difficulties during intra-operative navigation. 2D fluoroscopy is limited to projection imaging necessitating the collection of multiple scans from different directions. 3D fluoroscopy units overcome the drawbacks of the 2D counterpart, however, they are not cost-effective and currently are not widely available. CT-guided procedures still require the use of intra-operative 2D fluoroscopy. Moreover, both of these modalities operate with ionizing radiation which causes severe safety concerns to both surgical team and the patient. In order to provide a solution to the radiation exposure and improve the navigation accuracy, ultrasound (US) has been proposed by various research groups as an alternative to 2D fluoroscopy, as it provides real-time, non-

radiation based 2D/3D imaging. US current limitations include low signal-to-noise ratio (SNR), imaging artifacts, limited field of view and being a user operated imaging modality. all of these factors have hindered the widespread use of US in CAOS procedures.

The goal of this research is to develop robust, accurate, real-time and automatic image segmentation and localization methods for bone structures in ultrasound (US)-guided interventional procedures. A correct segmentation of bone structures is crucial in US-guided orthopedic procedures. In clinical practice, the quantification of these structures is generally performed by manual tracing, which is a time consuming and introduces inter- and intra-user variability errors. Hence, reliable, rapid, accurate, and automatic or semiautomatic methods of bone segmentation are required.

The focus of this work is to develop automatic segmentation techniques that are robust and accurate for ultrasound (US) guided minimally invasive surgeries. This research is intended to advance the larger goal of developing a novel US based computer assisted orthopaedic surgery (CAOS) system for minimally invasive bone reduction procedures. Such a system will ultimately address a variety of problems with the planning and execution of orthopaedic surgery procedures. Specifically we have investigated the potential and feasibility in using deep learning-based methods for real-time segmentation of bone surfaces. The goals of this research include the following:

- Developing new and robust deep learning-based methods that can allow automatic and real-time extraction of bone surfaces and surgical tools from two dimensional (2D) US scans with sufficient accuracy.

- Introducing multimodal convolutional neural network(CNN)-based technique for segmenting bone surfaces from in vivo US scans, in which fusion of feature maps and multimodal images are incorporated to abate sensitivity to variations that are caused by imaging artifacts and low intensity bone boundaries.

- Developing a block-based CNN for segmentation of bone surfaces from in vivo US

scans. The novelty of our proposed design is that it utilizes fusion of feature maps and employs multi-modal images to abate sensitivity to variations caused by imaging artifacts and low intensity bone boundaries.

- Developing a conditional Generative Adversarial Network (cGAN)-based method for accurate real-time segmentation of bone shadow regions from in vivo US scans. A novel GAN architecture designed to perform accurate, robust and real-time segmentation of bone shadow images from in vivo US data, is proposed. We show how the segmented bone shadow regions can be used as an additional proxy to improve bone surface segmentation results of a multi-feature guided (CNN) architecture.

- Developing a GAN-based computational method in order to produce synthetic B-mode bone US images and to generate their corresponding segmented bone surfaces which can be used as labels. Our model utilized self-projection and self-attention blocks in its architecture and attempts to provide a solution to two of the medical data main problems: scarcity of data size, due to a lack of standardized data, and patients' privacy concerns. We show how such an approach can improve bone surface segmentation accuracy using synthesized B-mode bone US images generated by this model when tested using a segmentation CNN.

1.2 Basics of Ultrasound Imaging

Ultrasound waves are non-audible sound waves that can be described as pressure waves of mechanical energy that is transmitted through a medium by vibration of molecules [4]. Sound is characterized by its frequency in hertz (Hz), wavelength in millimeters (mm) and amplitude in decibels (dB). Sound is a longitudinal wave. Air molecules are concentrated at the peak and there is refraction in the trough. As for most vibrations, there is a frequency of compression and rarefaction. Audible signal has a frequency of less than $20kHz$. Ultrasound has a greater frequency than this and in medical applications, the frequency varies

from 1 MHz to $15MHz$. The time for a sound wave to complete a continuous cycle is the time period; the measurement length is the microseconds. Wavelength is the amount of space on which the cycle takes place; it is directly proportional to the time from the start to the end of the process. Frequency relates to the quantity of repeated cycles per second measured in hertz (Hz). Acoustic speed is the rate at which the sound wave travels through the channel. It is equivalent to the frequency period of the wavelengths. Speed c is calculated by the stiffness and density of medium as:

$$c = \sqrt{K/\rho} \tag{1.1}$$

where ρ is the density is the density of the medium and K is the adiabatic elastic bulk modulus. The stiffness is the resilience of the component to pressure. The rate of propagation increases with increasing in stiffness and a decline in density. The ultrasonic probe is a very critical sensor that produces acoustic signals and also senses the returned signals. The efficiency and image quality of the ultrasonic scanner are greatly influenced by the characteristics and configuration of the probe (piezoelectric material, matching layer and acoustic lens). The transducer is a major component of the ultrasound system. The transducer probe produces a pressure wave and generates an echo. It is the mouth and ears of the ultrasound system. The transducer probe creates and captures waves of sound using a principle called the piezoelectric (pressure) effect. There are either one more than two quartz crystals in the sample or piezoelectric crystals. If electrical field is applied to such crystal, they quickly change structure.

Ultrasound is demonstrated by a variety of properties as it moves through the tissue on of which is reflection. Wave is reflected on boundary surfaces among tissue with various wave propagation property (such as fat , muscle, and blood). The level of reflection depends on the degree of variance. The leftover ultrasound energy will either infiltrate further or be absorbed by the tissue. Usually, the ultrasound wave strikes a variety of different reflector surfaces. The amplitudes of ultrasonic reflections rely on the standard acoustic impedance

Z of the neighboring tissues. The characteristic impedance of a tissue is equivalent to the result of its density, ρ, and velocity of propagation:

$$Z = c\rho \tag{1.2}$$

Soft tissue has an average speed of 1540 ms^{-1}. Reflection can be called either specular or diffuse. Specular reflection happens as sound waves meet large smooth objects, such as bones, which result in sound waves being reflected back in a relatively uniform direction. The cells in most soft tissue establish a more complex pattern in reflection on the transducer. Echo production is critically responsible for the presence of an acoustic impedance difference between the two classes of tissues. Acoustic impedance is the function of the tissues and is described as the product of a volume of the tissues and the speed at which the sound waves spread through the surface. If two tissues forms have same acoustic impedance, no resonance will be generated, as no waves of sounds will be reflected.

Non-specular reflections arise where the boundary is small than one wavelength of the ultrasound beams. The reflecting beam produced is called an Echo. Diffuse reflection is the dispersion of light that happens as it is reflected off the surface. Unlike the specular reflection, which is measured on the basis of the surface angle, the diffuse reflection is determined on the basis of the surface structure itself. For example, a rough surface absorbs light at several angles, depending on its bumps, pivots, and grain. Although a very smooth surface, owing to the molecular structure of the material, creates a distorted reflection at certain angles. Most acoustic radiation is absorbed and converted into heat as the sound waves travel through the skin. Instead of low frequency waves, high frequency sound waves are consumed. Variable frequency sound waves have a higher resolution than low-frequency sound waves, which is done at the disadvantage of reduced penetration. The inability to sufficiently penetrate the tissues of high-frequency sound waves results immediately in increased amplification and acoustic energy transfer as heat. The absorbance rate which occurs in the medium itself is also a factor in which some of the materials are more

diminishing than others. The total attenuation by a certain channel shall be calculated by the decibel coefficient per cm per MHz. Refraction happens as sound waves travel at varying transmission speeds from one origin to the next. This shift in speed leads to refraction or a shift in the direction of the main wave of sound. It is a change in the direction of a beam of Ultrasound at a boundary between two media. Refraction thus results in a weakening of the signal propagated. The traditional settings for the ultrasound instrumentation measure the return waves as if they were moving in a straight line. This results in a loss of image clarity as the refraction increases. The relationship between the velocity properties of the various tissues also affects the direction of the refraction. If the propagating sound wave is quicker in the first tissue due to less tissue impedance, the refraction would be more perpendicular. If the impedance is lower in the second tissue, resulting in higher sound wave propagation, the refraction happens away from the initial direction. Figure 1.1 depicts this an example of sound wave propagation in tissue, where incident, reflected and transmitted waves are indicated [5].

1.3 Deep Learning-Based Segmentation for Medical Images

Artificial neural networks are machine learning techniques that attempt to simulate and mimic the biological nervous system's mechanism of learning. The general objective is to implement and train mathematical models to produce specific desired outcomes. The idea is that machines may learn to perform tasks from experience provided in the form of training data. A neural network computes a function of the input data by propagating the computed values from the input neurons to the output neurons and using the weights as intermediate parameters. More specifically, learning occurs by changing the weights connecting the neurons [6]. Machine learning is a sub-field of artificial intelligence, under which learning methods can be divided into three main categories: supervised learning, unsupervised learning and reinforcement learning. Generally, machine learning models are utilized for supervised learning, in which the computer is given a set of labelled data and tasked with

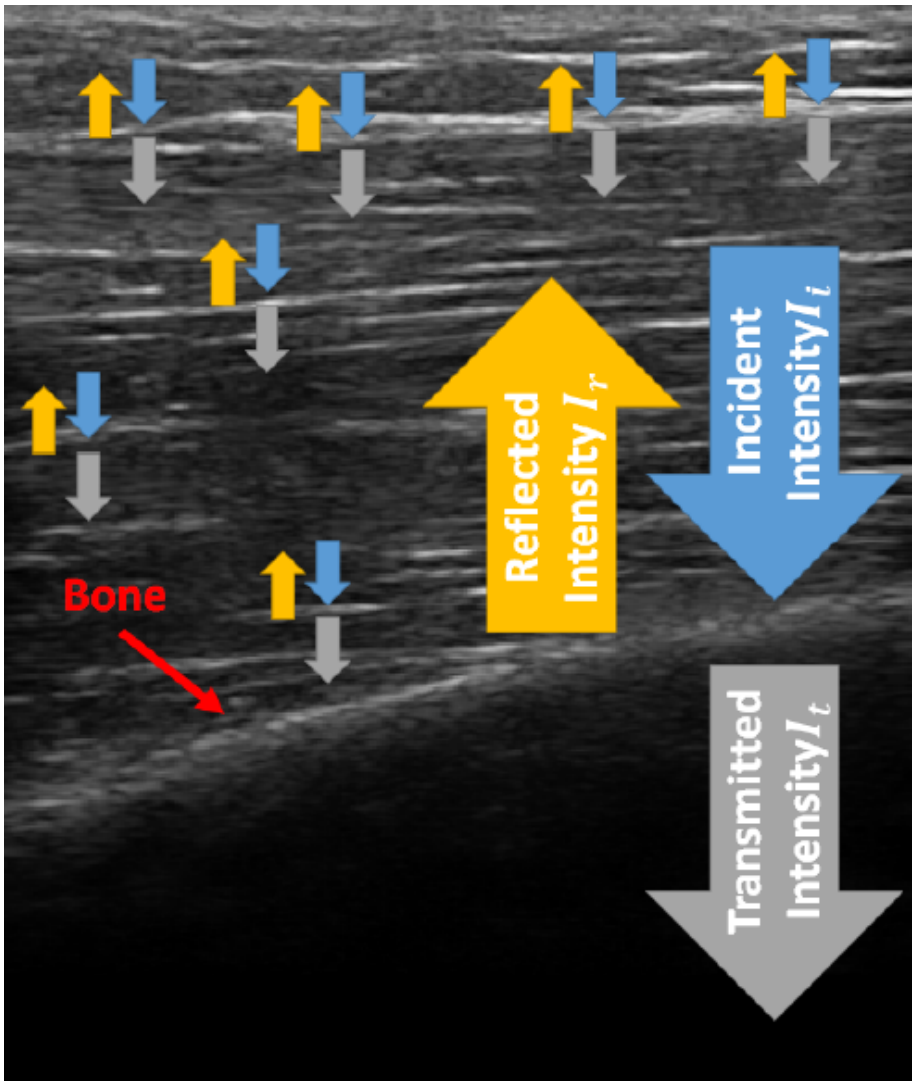

Figure 1.1: An example of sound wave propagation in tissue. Incident, reflected and transmitted waves are respectively indicated as blue, orange and gray arrows.

generating correct labels previously unseen data. In other words, supervised learning adjusts network parameters by a direct comparison between the actual network output and the desired output. Supervised learning is a closed-loop feedback system, where the error is the feedback signal. The error measure, which shows the difference between the network output and the output from the training samples, is used to guide the learning process. The error measure is usually defined by the mean squared error (MSE). In contrast to supervised learning, unsupervised learning does not require labeled data. Instead, it aims to recognize patterns in the data. Therefore, unsupervised learning involves no target values. It tries to associate information from the inputs with an intrinsic reduction of data dimensionality. Unsupervised learning is solely based on the correlations among the input data, and is used to find the significant patterns or features in the input data without the help of a teacher. Unsupervised learning is particularly suitable for biological learning in that it does not rely on a teacher and it uses intuitive primitives like neural competition and cooperation cite . A common example of unsupervised learning is clustering algorithms, which take a large set of data points and find groups within them. Reinforcement learning is similar to how a human learns from a trial and error experiment. It first decides on a certain action and then observes data to determine its effect. Over time, the model learns the best way to react. The use of machine learning has been increasing rapidly in the medical imaging field, including computer-aided detection, automated diagnosis and analysis. In 1995, Vapnik proposed a support vector machine [7] which became one of the most popular classifier at the time, i.e. the widespread use of SVM classifiers and clustering algorithms such as k-nearest neighbor (k-NN). These learning models are based on handcrafted features, i.e. manually extracted features from raw data or features extracted by other models. Also that year, random forests were propsed by Ho et al. [8].

The term deep learning was introduced by Hinton et al. in 2007 [9]. Deep learning has been an occurring major trend in health-related research areas, in the past decade, from medical informatics and bioinformatics, to sensing and medical imaging, if the increasing

number of publications is any indication [10]. In medical imaging, CNN-based approaches, in particular, have proven to be robust and adept in computer-aided segmentation, detection, and shape analysis [10, 11]. Pixel-level labeling in semantic segmentation was addressed by Long et al. [12] in the Fully Convolutional Network (FCN) design, which enabled the network to learn appropriate feature representations, e.g. pixel labeling. However, the FCN [12] suffers from limitations in labeling resolution due to its fixed-size receptive field and an overly simple deconvolution procedure, which Noh et al. in [13] have ameliorated by introducing an extension, composed of deconvolutions and unpooling layers, to the original network. A similar encoder-decoder network architecture was introduced by Badrinarayanan et al. [14] in which the encoder network is identical to [15]. However, the decoder upsamples low-resolution input feature maps using pooling indices, computed in the max-pooling step of the corresponding encoder, to perform non-linear upsampling. The encoder is typically a classification network consisting of convolutional layers that maps the input to a low resolution representation. The decoder, on the other hand, maps the low resolution feature maps to upsampled segmentation outputs.

Based also on FCN [12], Ronneberger et al. [16] developed a CNN that has been adapted towards segmentation of touching cells in microscopy. In this u-shaped encoder-decoder network architecture, more feature channels were added to allow the network to propagate context information to higher resolution layers. Furthermore, this network utilized a tiling strategy, that extrapolated the context information of the border pixels by mirroring the input image. In Addition, the problem of insufficient dataset size was addressed by utilizing data augmentation, i.e. by applying elastic deformation to the dataset, mimicking realistic deformation in cells. While, the architecture proposed in [16] has satisfactory performance results on various biomedical imaging tasks, it has two drawbacks; namely, its inability to fuse extracted feature maps, and its incapability to utilize multiple imaging modalities, which abates the segmentation accuracy.

Hazirbas et al. [17] investigated the fusion of feature maps using two different architec-

tural designs, namely sparse and dense. It was shown that utilizing depth information, for example, in addition to the appearance information, would improve the semantic segmentation performance. Furthermore, the work done by Valada et al. [18] proposed a semantic segmentation architecture based on a fusion technique that qualifies the learning of features from multiple imaging modalities. The end-to-end semantic segmentation architecture in [18] consists of experts that map imaging modalities to output segmentations, a Convoluted Mixture of Deep Experts (CMoDE), and a fusion layer that further learns corresponding fused kernels. One advantage of using multimodal images is that it is minimally sensitive to variations caused by noise and other external conditions. The Mixture of Experts (MoE) model, was originally introduced by Hinton et al. [19], where the input is mapped to the output by experts, over which a probability distribution is produced by a gating network to reduce the computational cost of multiple experts. The aforementioned experimental techniques were considered in our proposed design.

1.4 Basics of Convolutional Neural Networks

Convolutional neural networks (CNNs) were first introduced in 1980 by Fukushima's Neocognitron work [20], in which the receptive field of a convolutional unit with a given weight vector is shifted step by step across a two-dimensional array of input values. Neocognitron is similar to the architecture of supervised, feedforward deep learners with alternating convolutional and downsampling layers. However, a more practical revival of CNNs occurred by AlexNet [21] in the ImageNet competition in 2012. Much of its success might be attributed to the dropout training technique and the use of GPUs, ReLU function, and the unique techniques for generating more training examples by deforming the existing ones [22]. It is well known that a deeper network may better approximate the target function with increased nonlinearity and obtain better feature representations. However, with increased network complexity, training becomes more difficult and many issues might occur, e.g. overfitting (according to the bias–variance trade-off), vanishing gradient, and compu-

tational load. Deep CNNs provide a solution to these difficulties. A typical CNN consists of the following layers.

Convolutional layers: A convolutional layer detects local conjunctions of features from the previous layer. Neurons in a convolutional layer are organized in feature maps, within which each neuron is connected to local patches in the feature maps of the previous layer through a set of weights called a filter bank. Such a neighborhood is the neuron's receptive field in the previous layer. A convolution layer is composed of several convolution kernels, which are used to compute different feature maps. The convolution operation involves sliding a small filter (kernel) of size $K \times K \times n_c$ over the input $W \times H \times n_c$ corresponding to activations from the previous layer. The depth n_c of the kernel corresponds to the depth of the input. At each receptive field, the pixel values in the input are multiplied with pixel values in the filter in an element-wise manner and summed. This process makes the filter becomes a feature identifier and results in the learning of object features. Weights are learned during the training process and shared for computation in each layer. The stride, S, represents the increments by which the filter is moved. Stride length denotes the gap between two subsequent filter application locations, which reduces the size of the output tensor. For example, consider an input with size $32 \times 32 \times 3$ and a 5×5 filter would result in 75 weights in addition to the bias parameter [23]. Zero padding of dimension P might be applied to help capture features along the edges during convolution. This results in reducing the width and height of the output. The output's width W_o and height H_o are given by:

$$W_o = \frac{W - K + 2P}{S} + 1, H_o = \frac{H - K + 2P}{S} + 1 \tag{1.3}$$

Notice that the output's depth corresponds to the number of filters used, i.e. size $W_o \times H_o \times n_c$.

Activation layers: A post-processing layer applied right after the convolution layer is referred to as an activation layer. Convolutional layers' tensors are fed through non-

linear activation functions which facilitate learning of complex mappings between inputs and outputs. The most common activation function is rectified linear unit (ReLU) which computes the function $f(x) = \max(x, 0)$. One of the advantages of ReLU is that it accelerates the convergence of learning algorithms (e.g. stochastic gradient descent). Other non-linearities have been used before, e.g. the sigmoid function: $f(x) = \frac{1}{1+exp(-x)}$ and the $f(x) = \tanh(x)$ function. Leaky ReLU [24] and exponential linear unit (ELU) have also demonstrated an improved performances in some tasks.

Pooling layers: A pooling (or sub-sampling) layer often are utilized after a convolution layer. Pooling's role is to downsample the output of a convolution layer along spatial dimensions (height and width). For instance, a 2×2 pooling operation applied upon 12 feature maps will result in an output tensor with a $16 \times 16 \times 2$ size. The role of the pooling is to reduce the number of parameters learned by the network. Additionally, it will contribute in overfitting reduction and improve the accuracy of the network. Some of the most common pooling techniques include max pooling and average pooling. In a max pooling scenario, a feature map for each pooled area is replaced by a single maximum value amongst the others inside the pooled area. In an average pooling case, a feature map for each pooled area is replaced by a single value, which is the average of all values inside the pooled area. For an input volume $W \times H \times n_c$, the output volume's width W_o and height H_o are given by:

$$W_o = \frac{W - K}{S} + 1, H_o = \frac{H - K}{S} + 1 \tag{1.4}$$

Fully connected layers: A fully connected layer (also referred to as dense layer) is a final layer that all activations in the previous layer are connected to its neurons. Its output is usually the class score where the number of neurons is the same as the number of classes. One or more fully connected layers can be used in a CNN. The output of every fully connected layer is an $N \times 1$ vector.

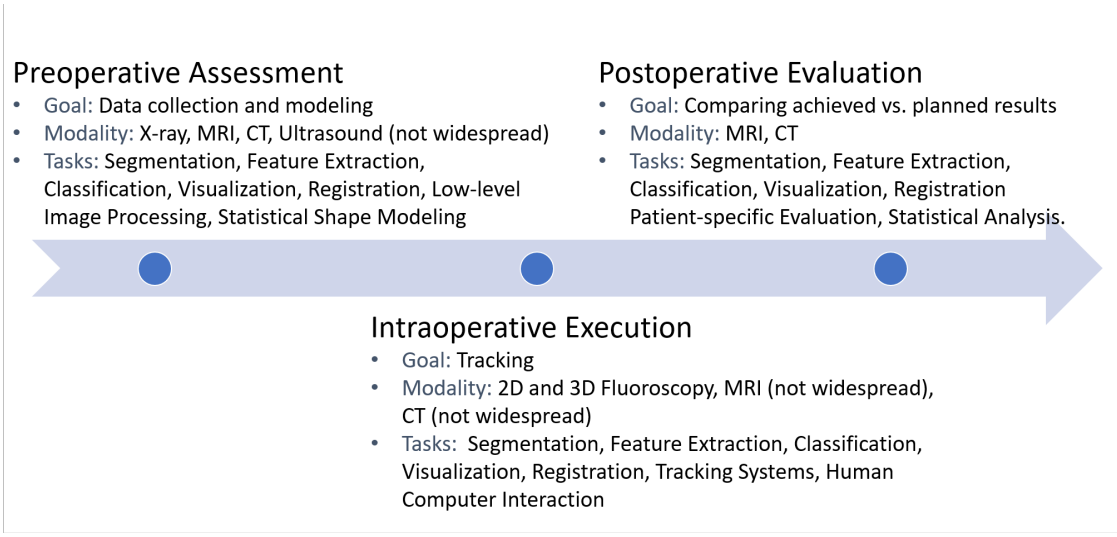

Figure 1.2: A typical three-phases CAOS workflow: Preoperative assessment, intraoperative execution, and postoperative evaluation.

1.5 Computer-Assisted Orthopaedic Surgery

A wide range of computer-based technologies that aim to improve orthopaedic surgical procedures is commonly referred to as computer-assisted orthopaedic surgery (CAOS). Many advancements in the fields of medical imaging and spatial tracking have contributed in improving the accuracy of navigation during an orthopaedic surgery. CAOS-based procedures are not identical and their framework may consist of different components due to the availability of medical devices and the application at hand. The three phases of a CAOS system are preoperative planning, intra-operative execution, and post-operative evaluation. Typically, a US-based CAOS workflow would be as described in Figure 1.2.

As can be seen in Figure 1.2, imaging as a modality is present in all three phases. Theses modalities are 2D plain X-ray, computed tomography (CT), magnetic resonance imaging (MRI), and 2D/3D fluoroscopy [5]. The most commonly used imaging modality is 2D fluoroscopy for intraoperative visualization and guidance. The more expensive 3D fluoroscopy imaging modality has also been used for various CAOS procedures for intraoperative guidance. Although it showed improved surgical outcome compared to its 2D

counterpart, its high cost has hindered its use. Moreover, the ionizing radiation exposure from fluoroscopy and CT modalities remains a major concern for the safety of patients and the surgical team. When a patient is admitted because of a fractured pelvic injury, a CT scan is the conventional protocol to follow. As [25] details in his work, the CT scan is taken to assess the fracture preoperatively and it will serve as a valuable reference model which could aid the surgical procedure. This allows the quantitative surgical plan, including the desired reduction of bone fragments and the ideal screw insertion locations. A challenge arises when the surgical team tries to transform the preoperative model into an intraoperative reference (US-based) as it requires direct knowledge of the patient's anatomy [26, 27, 25]. For an intraoperative phase in a CAOS procedure, it is essential to be able to precisely locate the patient's bone structures, i.e. bone segmentation. This way one can perform image registration to align intraoperative image to the preopertaive one. If a US-to-CT registration is achieved, one can visualize the bone structure of the patient in real-time fashion. There are many medical benefits to such a process including: (1) improving the surgical team's ability to accurately decide where to insert the surgical tools (e.g. screw placement), (2) reducing the ionizing radiation, and (3) improving the postoperative procedure.

1.6 Challenges in Ultrasound Bone Segmentation

While ultrasound (US), as an imaging modality, provides a low cost, safe ionizing radiation-free alternative to conventional fluoroscopy in CAOS surgeries, it suffers from major disadvantages in its current technology state. US has a low SNR ratio which makes scans of bone structures appear as soft tissues, i.e. muscles. In addition, an abundance of speckle presence degrades the quality of the US image and differentiating anatomical boundaries more challenging [28]. Some US scanners are typically operated by two individuals and the fact that the probe is manually-operated introduces changes in orientation with respect to the bone surface. This changing orientation can alter the appearance significantly [29]. As an active research are, many have investigated segmentation of bone structures for dif-

ferent applications. Thomas et al. [30] studied bone segmentation for the task of aiding fetal femur length calculation. Hacihaliloglu et al. [31] developed Phase Symmetry (PS) US bone segmentation as a precursor for US-CT registration [32]. In their work and sing Log-Gabor filters over multiple scales and orientations, phase symmetric regions were extracted in 2D and 3D. A more optimized automated method was also introduced by [33, 34] achieving a mean surface fit error of 0.62 mm when tested on clinical data [25]. Jain et al. and Foroughi et al. [35, 36, 37, 29, 38, 39] have demonstrated how to improve bone segmentation by utilizing shadow regions in a US bone scan. Foroughi et al. and Hacihaliloglu et al. [36, 40] have used dynamic programming methods to improve intensity and phase-based bone segmentation by minimizing disconnected surfaces. Deep learning-based methods have also been put to the task of bone segmentation from US bone images. Salehi et al. [41] proposed a U-net [16] based segmentation method of bone surfaces from US data, in which recall and precision rates were reported at 0.87. Similarly, Baka et al. [42] proposed another U-net based segmentation method of bone surfaces from US data in which the recall rate was 0.94 and the precision rate was reported at 0.88. Villa et al. [43] proposed a fully convolutional network (FCN) [12] based segmentation method of bone surfaces from US data, in which the recall rate was 0.62 and the precision rate was reported at 0.64. However, in these studies, bone surface localization accuracy was not investigated and low-quality bone surfaces were excluded from the validation and testing procedures. In addition, none of these models were designed specifically for the task of bone segmentation. As can be expected, networks whose architectures were designed to extract features from US data could improve the accuracy of the desired outcome.

CHAPTER 2
BONE SURFACE SEGMENTATION FROM ULTRASOUND USING MULTIMODAL CNN

In this chapter, we present our development of a multimodal convolutional neural network (CNN)-based technique for segmenting bone surfaces from in vivo US scans. The proposed design utilizes fusion of feature maps and multimodal images to abate sensitivity to variations that are caused by imaging artifacts and low intensity bone boundaries. In particular, our multimodal inputs consist of B-mode US images and their corresponding local phase filtered counterparts. Various fusion operations were investigated for our proposed network using different fusion architectures. Both quantitative and qualitative evaluations were performed on our designs in order to demonstrate statistically significant performance compared to others' state-of-the-art networks. Promising results, which showed accurate and robust segmentation of bone surfaces, were observed and further validations would be required prior to utilizing the proposed methods clinically.

2.1 Introduction

The annual rate of different bone-related musculoskeletal disorders (MSDs) has significantly increased during the past decades [2]. This includes chronic low back pain, spinal fracture, cervical spinal stenosis, spinal osteoarthritis, and other MSDs conditions. These conditions would cause pain and affect body functionality and could require spinal surgery such as spinal fusion, Laminectomy, Foraminotomy, Discectomy, etc. Additionally, spine related injuries for the years 2002-2004 were estimated at $193.9 billion [1].

The use of medical imaging in the preoperative, perioperative and postoperative periods during the surgical procedure is essential to determine the surgical procedure and treatment progress. Imaging is one of the most important components of any computer assisted ortho-

pedic surgery (CAOS) system. The standard intra-operative imaging modality in CAOS is 2D/3D fluoroscopy. Both of these modalities operate with ionizing radiation which causes important safety concerns to both surgical team and the patient. A safe intra-operative imaging alternative, ultrasound (US) has been incorporated into various CAOS systems. US provides real-time, non-radiation based 2D/3D imaging. However, low signal-to-noise ratio (SNR) and imaging artifacts have hindered the widespread use of US in CAOS procedures. In order to provide a solution to these difficulties, focus has been given to develop automated US bone segmentation and enhancement methods that are robust and computationally inexpensive for US guided CAOS procedures.

Machine learning methods have been recently proposed for the segmentation of bone surfaces from US data. Ozdemir et al. [44] investigated the use of a supervised learning framework that combines the physical constraints of US into a graphical model for segmentation. Although low quality bone surfaces were excluded from the validation and testing procedure, in [42], a network architecture based on U-net of [16] was investigated for segmenting vertebra bone surfaces. Reported precision and recall rates were 0.88 and 0.94, respectively. Also based on [16], a deep learning network architecture was developed by [41] for segmentation of bone surfaces from US data. Although localization accuracy was not reported, the recall and precision rates for the proposed method were 0.87. In [43], an algorithm based on fully convolutional networks (FCN) was proposed for automatic localization of the bone interface in US images and the results were compared against those of confidence in phase symmetry (CPS).

2.2 Methods

2.2.1 Data Acquisition

A total of 261 B- mode US images, from twelve healthy subjects, were collected using Sonix-Touch US machine (Analogic Corporation, Peabody, MA, USA), after obtaining the institutional review board (IRB) approval. Depth settings and image resolutions varied

between 3-8 cm, and 0.12-0.19 mm, respectively. Data augmentation (by means of image rotation) was performed on this dataset to obtain 1,044 B-mode US images in total. All bone surfaces were manually segmented by an expert ultrasonographer.

2.2.2 Local phase image features

A local phase image is a combination of, in this case, three different image filters which gives an enhanced and robust view of a desired 2D/3D ultrasound (US) data. It is based on the use of a gradient energy tensor (GET) and is modified to be a new feature enhancement technique. In this specific scenario, it allows us to enhance bone features from 2D/3D US data. The local phase image feature extraction is based on the computation of three different phase image features [45]. The combination of three different image phase features provides a more compact and robust enhancement. Next we explain how these three features are extracted and combined.

Hacihaliloglu et al. [46], proposed a tensor-based phase feature descriptor called local phase tensor image feature ($LPT(x, y)$). $LPT(x, y)$ is obtained from B-mode US image, $US(x, y)$, using even (T_{even}) and odd filter responses (T_{odd}) which represent the symmetric and asymmetric features found in $US(x, y)$. T_{even} and T_{odd} filters are constructed using a gradient energy tensor (GET) filter [46]. The final $LPT(x, y)$ image is obtained as follows [46]:

$$LPT(x, y) = \sqrt{T_{even}^2 + T_{odd}^2} \times cos(\varphi). \tag{2.1}$$

Here ϕ represents instantaneous phase obtained from the symmetric and asymmetric feature responses, respectively [46]. $LPT(x, y)$ provides a general enhancement independent of the specific feature type present in the acquired $US(x, y)$ scans. This presents a more robust enhancement of complex shaped bone surfaces, such as the spine [46]. However, soft tissue interfaces close to the bone surface which have similar intensity values are also enhanced during this process. To suppress the enhancement of these soft tis-

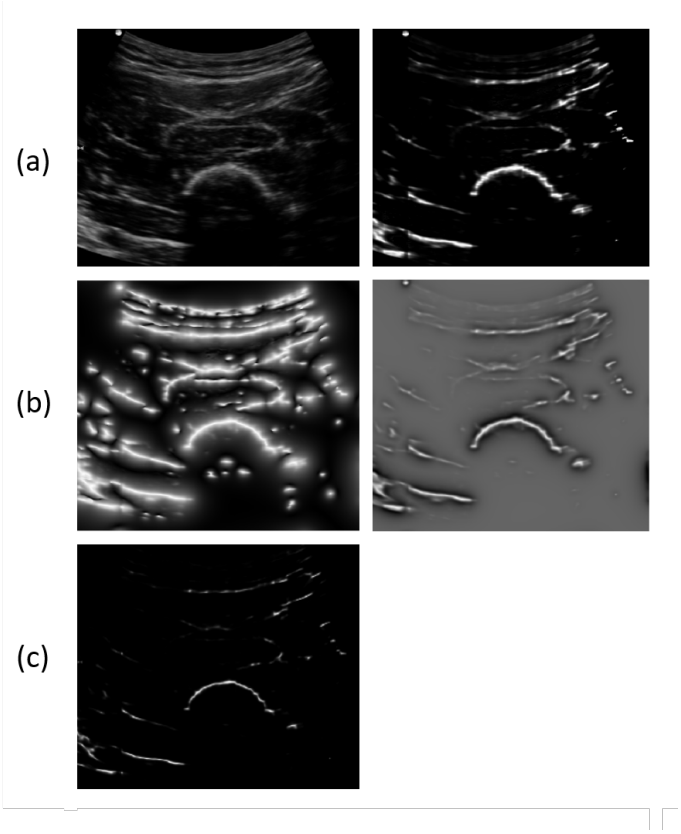

Figure 2.1: (a)-left: $US(x, y)$ image of in vivo femur bone. (a)-right: $LwPA(x, y)$ image. (b)-left:$LPT(x, y)$ image. (b)-right: $LPE(x, y)$ image. (c)-left: $LP(x, y)$ image.

sue interfaces and obtain a more compact bone representation, monogenic image filtering was applied to $LPT(x, y)$ image. This results in the extraction of two more local phase image features: local phase energy, $LPE(x, y)$, and local weighted mean phase angle, $LwPA(x, y)$. These two image phase features are obtained by combining the bandpass filtered $LPT(x, y)$ image, denoted as $LPT_B(x, y)$, with Riesz filtered components (represented by h_1 and h_2) resulting in the extraction of monogenic signal image, $US_M(x, y)$, as follows [45]:

$$\begin{aligned} US_M(x, y) &= [US_{M1}, US_{M2}, US_{M3}] \\ &= [LPT_B(x, y), LPT_B(x, y) * h_1, LPT_B(x, y) * h_2]. \end{aligned} \tag{2.2}$$

By accumulating the local energy of the image along multiple filter responses, the $LPE(x, y)$ image encodes the underlying shape of the bone boundary. Averaging the phase sum of the response vectors over many scales generates the $LPE(x, y)$ image as follows:

$$LPE(x, y) = \sum_{sc} |US_{M1}| - \sqrt{US_{M2}^2 + US_{M3}^2}. \tag{2.3}$$

Here, sc corresponds to the number of filter scales. The $LwPA(x, y)$ image can be found as follows:

$$LwPA(x, y) = arctan(\frac{\sum_{sc} US_{M1}}{\sqrt{\sum_{sc} US_{M1}^2 + \sum_{sc} US_{M2}^2}}). \tag{2.4}$$

The $LwPA(x, y)$ image preserves all the structural details of the $LPT(x, y)$ image, i.e. soft tissue interfaces and bone surfaces. Investigating the extracted local phase images ($LPT(x, y)$, $LPE(x, y)$, $LwPA(x, y)$) in Figure 2.1 we can see that the bone surfaces have accurate localization in all the extracted phase images. However, the soft tissue interfaces do not have similar localization accuracy (they appear in different local regions in the image). Using this investigation the final local phase bone image, $LP(x, y)$, is obtained by multiplying the three phase feature images as: $LP(x, y) = LPT(x, y) \times LPE(x, y) \times$

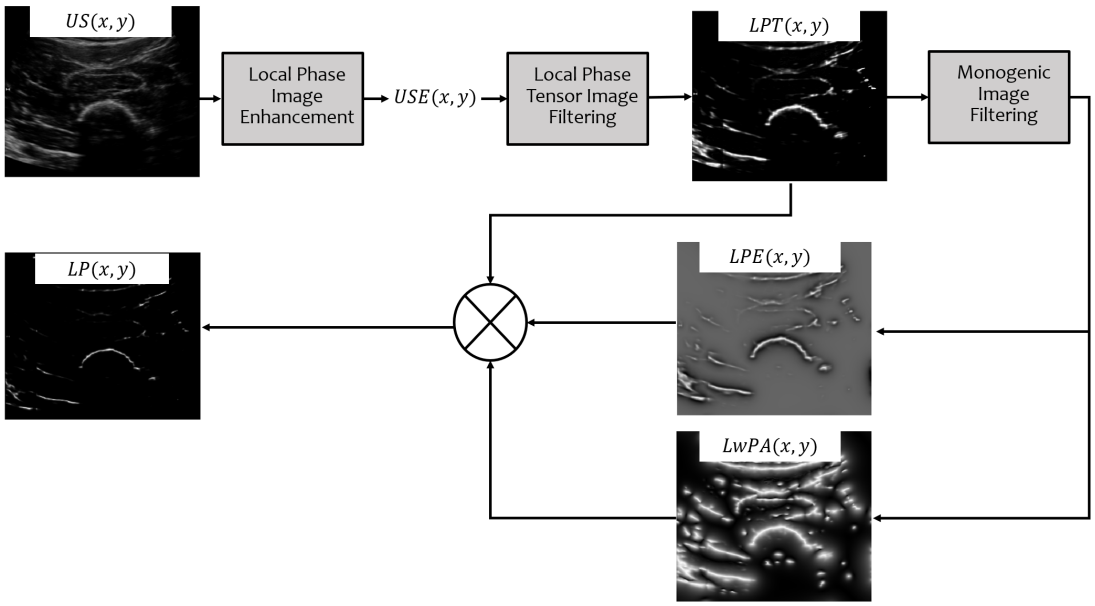

Figure 2.2: Flowchart summarizing the process to obtain an $LP(x, y)$ image from $US(x, y)$ image.

$LwPA(x, y)$. These mathematical operations are pictorially depicted in Figure 2.2.

Figure 2.1 shows all the extracted three local phase image features and the final $LP(x, y)$ image. Investigating Figure 2.1 we can see that the final $LP(x, y)$ has compact representation of the bone surface with reduces soft tissue artifacts. The extracted local phase image, $LP(x, y)$, and the B-mode US image, $US(x, y)$, are used during the proposed CNN-based bone segmentation methods which is explained in the next section.

2.2.3 Fusion in Deep Learning

Fusion, as a tool for machine learning tasks, i.e. classification and segmentation, has been thoroughly examined by many including [47, 48, 49, 50]. Liu et al. in [48] inquired on several fusion models, i.e. early, halfway, and late fusion, based on a vanilla ConvNet to fuse feature maps from color and thermal input images. They introduce a 1×1 convolutional layer after their concatenation layer as a fusion operation. In early fusion, the concatenation of the feature maps from different branches at low-levels captures visual features, such as

corners and line segments. Similarly, the halfway fusion employs the aforementioned fusion operation after several convolutional layers to capture more semantic meanings while retaining some final visual details. In late fusion, concatenation occurs at the fully connected stage in order to execute high-level feature fusion. Liu et al. report that halfway fusion achieves the best synergy for their pedestrian detection application. Guo et al. in [47] evaluated medical image segmentation using the aforementioned fusion schemes to fuse multi-modal images, i.e. MRI, CT, and PET, to detect the presence of soft tissue sarcoma. They report that early halfway fusion has shown similarly good performance. However, the decision-level late fusion model performs the worst among all fusion schemes in their study.

Feichtenhofer et al. in [50] addressed the importance of spatial correspondence in channels when performing fusion in multi-stream networks. To demonstrate how the channels correspondence may vary when fusion is performed, Feichtenhofer et al. [50] list five ways of spatial fusion operations, as shown in Table 2.1. A fusion function f that fuses any two feature maps, $\mathbf{x}^a \in \mathbb{R}^{H\times W\times D}$ and $\mathbf{x}^b \in \mathbb{R}^{H\times W\times D'}$ can be defined as $f : \mathbf{x}^a, \mathbf{x}^b \rightarrow \mathbf{y}$, where $\mathbf{y}$ is the output map, where $\mathbf{y} \in \mathbb{R}^{H\times W\times D''}$.

Channel numbering in both sum and max fusions is arbitrarily assigned. Filters of each network may be optimized when the channel correspondence is properly employed. As can be concluded from Table 2.1, concatenation fusion does not define any correspondence as it stacks feature maps at the same spatial locations across the feature channels. In practice, subsequent layers define the correspondence by learning suitable filters that weight the layers. Convolution fusion is performed through a similar operation as the concatenation one, except that the stacked feature maps are convolved with a bank of filters and biases [50]. These four operations of fusion were considered in designing our model. Due to its high dimensionality, bilinear fusion was not considered. Based on the performances of the remaining four operations on a preliminary experiment, concatenation fusion was selected since it performed the best. Throughout our network designs, the used fusion operation is

Table 2.1: Spatial Fusion Operations

Type	Equation	Dimension
Sum	$y^{sum}_{i,j,d} = x^{a}_{i,j,d} + x^{b}_{i,j,d}$	$y^{sum} \in \mathbb{R}^{H \times W \times D}$
Max	$y^{max}_{i,j,d} = max\{x^{a}_{i,j,d}, x^{b}_{i,j,d}\}$	$y^{max} \in \mathbb{R}^{H \times W \times D}$
Concat	$y^{conc}_{i,j,2d} = x^{a}_{i,j,d}$, $y^{conc}_{i,j,2d-1} = x^{b}_{i,j,d}$	$y^{conc} \in \mathbb{R}^{H \times W \times 2D}$
Conv	$\mathbf{y}^{conv} = \mathbf{y}^{conc} * \mathrm{f} + b$	$y^{conv} \in \mathbb{R}^{H \times W \times D'}$
Bilinear	$\mathbf{y}^{bilin} = \sum_{i=1}^{H} \sum_{j=1}^{W} \mathbf{x}^{a\top}_{i,j} \mathbf{x}^{b}_{i,j}$	$y^{bilin} \in \mathbb{R}^{D^2}$

concatenation fusion.

2.2.4 Multimodal Fusion-Based CNN Architecture

We developed our proposed CNN architecture based on the common contractive-expansive design. First, we resize the input B-mode US image $US(x, y)$ and its complementary local phase filtered image $LP(x, y)$ based on [51] and [45] to a standardized 256×256 size. we considered three stages of fusion in our proposed designs; early, mid, and late fusions as depicted in Figures 2.3 through 2.5. Our early-fusion model, depicted in Figure 2.3, fuses the input B-mode US image, $US(x, y)$, and the local phase filtered image, $LP(x, y)$, at the pixel level. The fused image is then processed through a single network. In Figure 2.4, a feature level fusion model was implemented in which mid-level features from both primary and secondary networks are fused together. A 1×1 convolution is performed on the output of the fused layer. Finally, in Figure 2.5, a classifier level model was implemented in which high-level features from each network are concatenated. Again, a 1×1 convolution is performed on the output of the fused layer to generate the final segmented probability distribution. The performance of the proposed designs were compared against each other and the networks proposed by [16] and [17]. Prior to quantitative and qualitative validation, images were resized to their original size. Inspired by the network designs in [18] and

Table 2.2: Proposed CNN Layers Specifications

Layer	**Input**	**Output**	**Layer**	**Input**	**Output**
Input	$256 \times 256 \times 1$	$256 \times 256 \times 1$	Filter	$256 \times 256 \times 1$	$256 \times 256 \times 1$
Conv1	$256 \times 256 \times 1$	$256 \times 256 \times 32$	Conv2	$128 \times 128 \times 32$	$128 \times 128 \times 64$
Max1	$256 \times 256 \times 32$	$128 \times 128 \times 32$	Max2	$128 \times 128 \times 64$	$64 \times 64 \times 64$
Conv3	$64 \times 64 \times 64$	$64 \times 64 \times 64$	Conv4	$32 \times 32 \times 64$	$32 \times 32 \times 128$
Max3	$64 \times 64 \times 64$	$32 \times 32 \times 64$	Max4	$32 \times 32 \times 128$	$16 \times 16 \times 128$
Up1	$16 \times 16 \times 256$	$32 \times 32 \times 256$	Up2	$32 \times 32 \times 256$	$64 \times 64 \times 256$
Conv6	$32 \times 32 \times 256$	$32 \times 32 \times 256$	Conv7	$64 \times 64 \times 256$	$64 \times 64 \times 256$
Conc1	$32 \times 32 \times 384$	$32 \times 32 \times 384$	Conc2	$64 \times 64 \times 320$	$64 \times 64 \times 320$
Up3	$64 \times 64 \times 256$	$128 \times 128 \times 256$	Up4	$128 \times 128 \times 128$	$256 \times 256 \times 128$
Conv8	$128 \times 128 \times 256$	$128 \times 128 \times 128$	Conv9	$256 \times 256 \times 128$	$256 \times 256 \times 64$
Conc3	$128 \times 128 \times 192$	$128 \times 128 \times 192$	Conc4	$256 \times 256 \times 96$	$256 \times 256 \times 96$

[17], in our proposed design, each input image would connect to an independent primary network, and a secondary network, as depicted in Figure 2.5. In each network, the input image is convolved in the encoder by convolutional layers with 3×3 filters (same padding convolutions) each followed by a rectified linear unit (ReLU) and a 2×2 maxpooling. Whereas in the decoder path, transposed-convolutions of same kernel size and paddings are applied and upsampled. The encoder maps the input image into a low-dimension latent space, and the decoder maps the latent representation into the original space. The proposed network layers specifications are tabulated in Table 2.2. In the primary network, the input image is a B-mode US image US(x,y), while in the secondary network, the input is a local phase filtered image LP(x,y) that proceeds through the aforementioned convolutional, and max pooling layers. Feature maps extracted from both networks are fused in a late fusion stage. This classifier level model was implemented in which high-level features from each network are concatenated. A 3×3 convolution with sigmoid activation is performed on the output of the fused layer to generate the final segmented probability distribution [17].

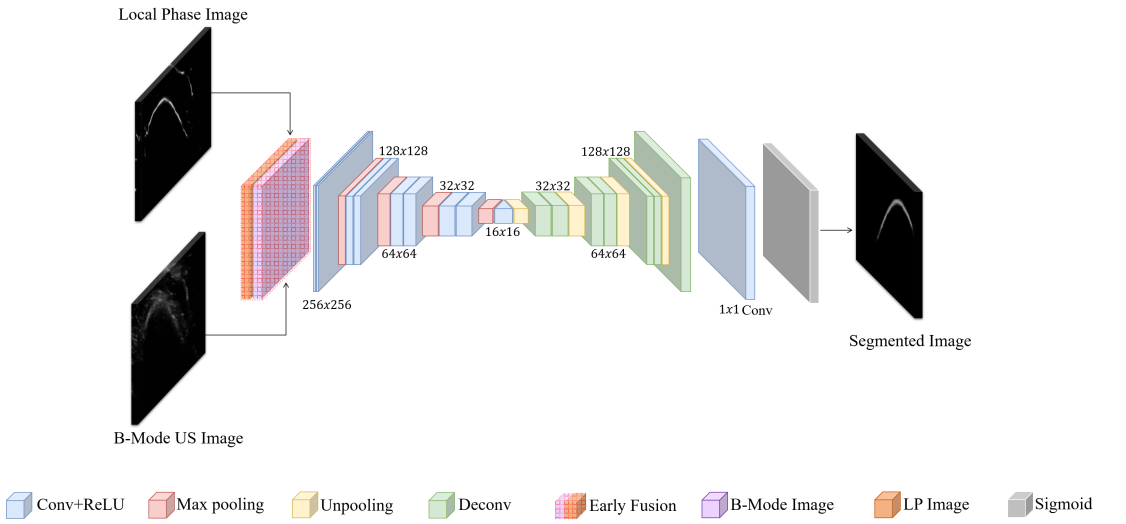

Figure 2.3: An overview of the early-fusion CNN architecture. Input B-mode US image, $US(x, y)$, is concatenated with the local phase filtered image, $LP(x, y)$, at the pixel level, and the result is processed through the network.

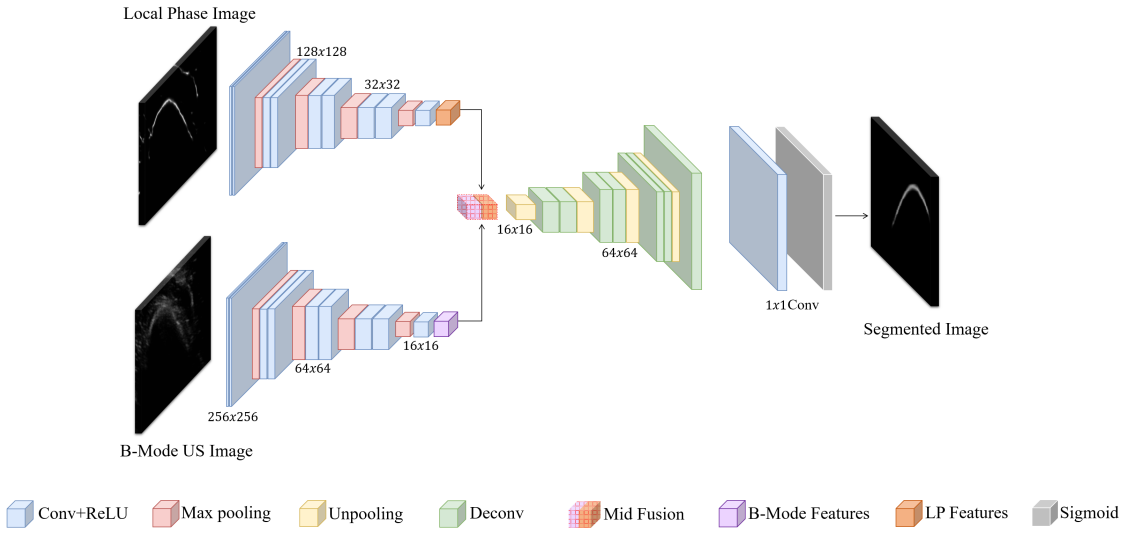

Figure 2.4: An overview of the mid-fusion CNN architecture. Input B-mode US image, $US(x, y)$, is processed through the primary encoder (bottom), while the local phase filtered image, $LP(x, y)$, is processed through the secondary encoder (top). Feature maps from both encoders are fused in a mid fusion stage, and processed through the decoder.

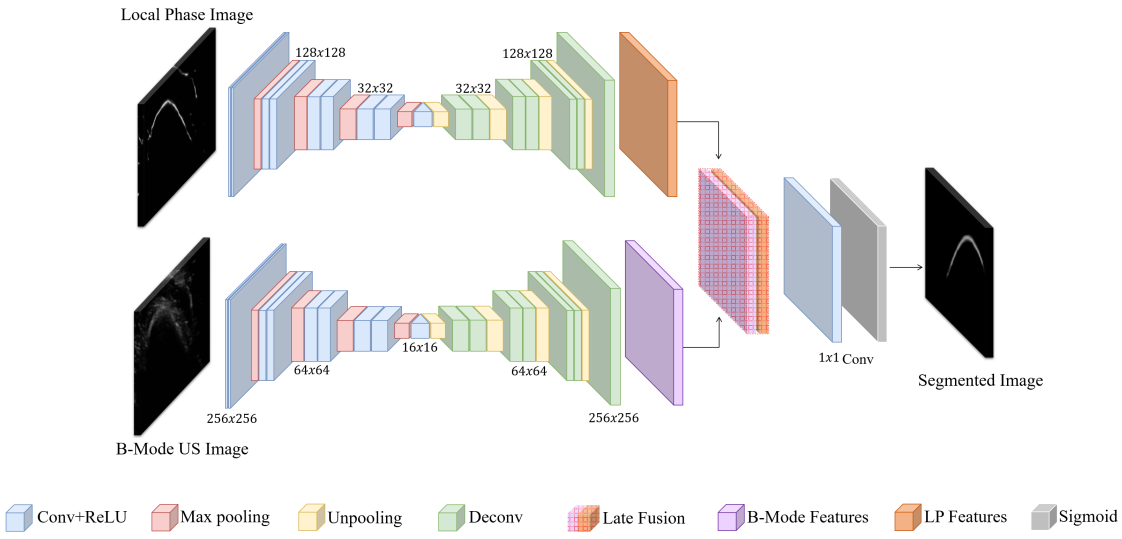

Figure 2.5: An overview of the Late-fusion CNN architecture. Input B-mode US image, $US(x, y)$, is processed through the primary network (bottom), while the local phase filtered image, $LP(x, y)$, is processed through the secondary network (top). Feature maps from both networks are fused in a late fusion stage.

2.2.5 Training and Testing

The performance of our proposed design was compared against the network proposed in [16] with its depth increased to a scale close to our proposed design. Our proposed design and U-net were trained using a training set of 912 B-Mode US images and their corresponding local phase filtered images. The remaining 132 B-mode US images were reserved for testing the performance of the networks. During the random split of the dataset, same scans were not used for both training and testing. This process was repeated five times, with each training and testing data randomized from our dataset. Both networks were trained to minimize the following cross-entropy loss function:

$$\mathcal{L}(X, Y) = -\frac{1}{n}\sum_{i=1}^{n} y^{(i)} \ln a(x^{(i)}) + (1 - y^{(i)}) \ln(1 - a(x^{(i)})) \tag{2.5}$$

where $X = \{x^{(1)}, ..., x^{(n)}\}$ denotes the training input images set, and $Y = \{y^{(1)}, ..., y^{(n)}\}$

denotes their corresponding mask labels set. Here, we employ a ReLU activation function to restrict $a(x)$ between $(0, x)$.

Based on [52], [40], [53] and [54], five error metrics were calculated in our testing set; namely, F-score, Rand error, Hamming Loss, as well as the IoU and average bone surface localization error. To measure how similar any two segmentation regions in an image with *k* pixels, P_1 and P_2 we denote the number of pixels in the same class, and the number of the pixels in the remaining class as *x*, and *y* respectively. Therefore, the Rand error, R_e, could be calculated using the Rand index, R_i, as:

$$R_e = 1 - R_i = \frac{x + y}{\binom{k}{2}} \tag{2.6}$$

In order to evaluate the accuracy of our segmentation, the pixel error between the manual segmentation and the B-mode US images is calculated by taking the squared Euclidean distance. Moreover, boundary labeling could be compared in the field of digital topology using the warping error [24]. This metric is used to measure the topological differences between structures. In particular, it can be used as a cost function in boundary detection since it tolerates disagreements over boundary location and penalizes topological disagreements. For a reference labeling L, we can find the warping error between L and some candidate labeling T as:

$$D(T||L) = \min_{L} ||T - L||^2 \tag{2.7}$$

It is important to note that while the warping error can penalize different topological errors, the Rand error penalizes only connectivity errors. For instance, in some medical imaging applications, the warping error does not penalize boundary location shifts. Therefore, the Rand error takes into account shifts in boundary locations. This is due to the fact that the warping error weights a topological error by the number of pixels involved in the error itself, while the Rand error weights a split or merger by the number of pixels in the

objects associated with the errors [53]. The bone localization error is calculated as the average Euclidean distance (AED) error between the automatically segmented bone surfaces and the manual expert segmentation.

2.3 Results

In this section, we discuss our results of the experiments that were carried out using the Keras framework and Tensorflow as backend with an Intel Xeon CPU at 3.00GHz and an Nvidia Titan-X GPU with 8GB of memory. On average, our networks converge in about 6 hours during the training process. We first demonstrate the performance of state-of-the-art network on our data set. We then show the results of our designs with different stages of fusion applied on our bone data. Furthermore, a comparison of error metrics calculated for four additional bone surfaces is included alongside its segmentation results.

2.3.1 Bone Segmentation Qualitative Results

Qualitative results of our late-fusion network design as well as the network in [16] are shown in Figure 2.6 (b) and (c), where the red pixels indicate high prediction scores while blue pixels indicate low prediction scores for the segmentation. The prediction outcome when only B-mode US images were used in training, as the case in Figure 2.6 (b) for U-net, had lower probability distribution compared to ours. The inadequate segmentation performance may be attributed to the nature of the US images used in the testing process. Figure 2.6 (c) shows an improved prediction outcome when both B-mode US and Local-phase filtered images are used to train our proposed network. In addition, bone localization results against expert manual localization, are presented in Figure 2.6 (d) and (e). Investigating the localization results we can infer that the output of U-net, trained only using B-mode data, had large gaps from the expert localization and missing bone boundaries. Qualitatively, our late-fusion design achieved better performance for this dataset.

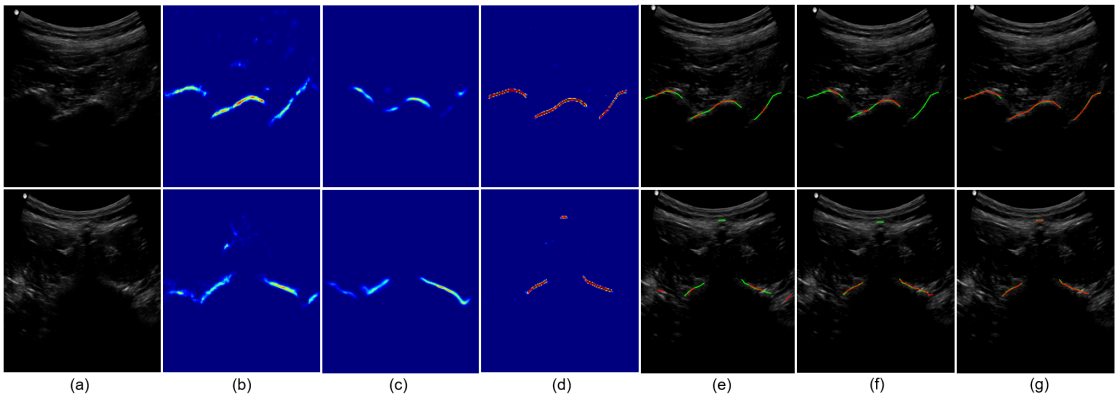

Figure 2.6: (a) Two US B-mode images of in vivo spinal bones, and their corresponding segmentation outputs from the following network architectures: (b) Ronneberger et al. [16] trained with B-mode US images only, (c) Ronneberger et al. [16] trained with B-mode US images and Local-phase filtered images, and (d) our proposed design trained with both B-mode US and Local-phase filtered images. Bone localization in (red) to manual expert localization (green) obtained from (e) no-fusion Ronneberger et al. [16], (f) early-fusion Ronneberger et al. and from (g) our late-fusion design.

Table 2.3: Bone Error Metrics

Method	IoU%	F-Score	Rand	Hamming	AED (mm)
Ronneberger [16] US B-mode only	0.803022	0.866023	0.662321	0.196977	2.9653
Ronneberger [16] US B-mode & LP	0.851270	0.936892	0.750253	0.062402	0.9372
Ours US B-mode & LP	**0.969489**	**0.977450**	**0.448840**	**0.030511**	**0.1097**

2.3.2 Bone Segmentation Quantitative Results

The aforementioned error metrics were calculated for both networks, and the results are tabulated in Table 2.3. As can be seen from Table 2.3, the average numerical error calculations show that that the late-fusion design had lower errors, and the highest average IoU and F-scores. A paired t-test, at a %5 significance level, between our designed network and the network proposed in [16] achieved p-values less than 0.05 indicating that the improvements of our method are statistically significant. U-net achieved overall AED error of 2.96 mm while our design achieved 0.1097 mm.

2.3.3 Accuracy Localization Study

To further validate our results, we have trained our proposed models and tested them on four more bone structures. AED error measurements obtained for the Radius (85 cases), Femur (16 cases), Knee (10 cases), Tibia (4 cases), bone surfaces are shown in Figure 2.7. The overall AED for our late-fusion design is 0.1255 with a standard deviation of 0.006. This represents a significant improvement compared to the other networks.

2.3.4 Fusion Comparison Study

Results for four additional bone surfaces of our early, mid, and late-fusion network designs were compared against state-of-the-art networks as shown in Figure 2.8, where the red pixels indicate high prediction scores while blue pixels indicate low prediction scores for the segmentation. The prediction outcome of the network in [16] displayed in Figure 2.8 (b) had the lowest probability distribution amongst all others. The inadequate segmentation performance may be attributed to the nature of the US images used in testing process. For low quality US scans, where the bone surface has a low intensity profile and high intensity soft tissue interfaces appearing above the bone surface, the performance of the network proposed in [16] decreases. The importance of collecting high quality US data and its affect on the segmentation outcome was also discussed previously in [42] who proposed a similar network architecture for segmenting vertebra bone surfaces from US data. The segmentation results of the network in [17] and mid-fusion network in Figure 2.8 (c) and (d) respectively, are comparable. However, the network in [17] had a higher probability distribution than the mid-fusion network. This is because in the network proposed in [17], the fusion performed at the feature level is considered a slow fusion in which multiple feature maps are fused throughout the encoder. In contrast, fusion occurs only once at the last layer of the encoder in the mid-fusion design. On the other hand, as shown in Figure 2.8 (g), early fusion outperforms the network in [17] since the fusion happens at the pixel level in which the fused image would possess enhanced bone surfaces while the soft tissue

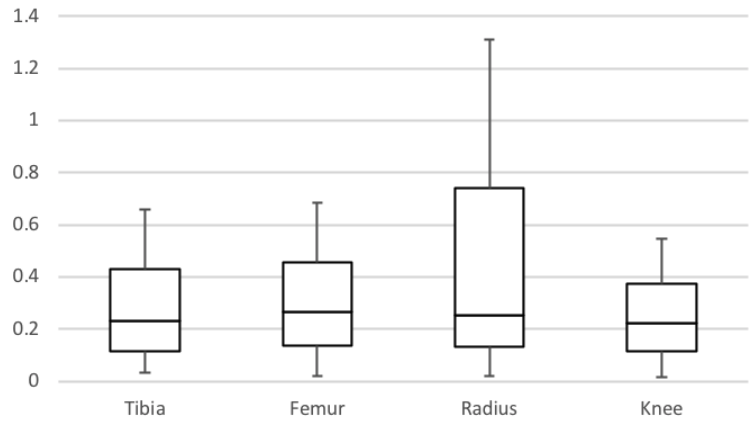

Figure 2.7: Accuracy Localization of Four Bone Types

interfaces remain unaltered. We also show the results of our single network when no fusion is performed; once when the network is trained with B-mode US images only, and once when the network is trained with local-phase filtered images only, as depicted in Figure 2.8 (e) and (f) respectively. Our late-fusion design outperformed all of the aforementioned models. Qualitatively, the late-fusion prediction results in Figure 2.8 (e) has no artifacts, no false prediction, and no fragmented segmentation.

For each network, a 5-fold cross validation was performed, with each training (300 images) and testing (115 images) data randomized from our dataset. The aforementioned error metrics were calculated for each of the networks, and the results are tabulated in Table 2.4. As can be seen in Table 2.4, the average numerical error calculations show that that the late-fusion design had the lowest errors in both warping and rand, and the highest average IoU and F-scores.

Qualitative results about the final segmented surfaces for the additional four bone structures and comparison against the gold standard segmentation (manual expert segmentation) are provided in Figure 2.9. As can be concluded, the segmentation accuracy is visibly improved when the local phase filtered feature maps are fused with those of the B-mode US images. As can be seen in Figure 2.9 (a), the segmentation results of the network in [16] suffered from false segmentation in some instances and fragmented segmentation in all

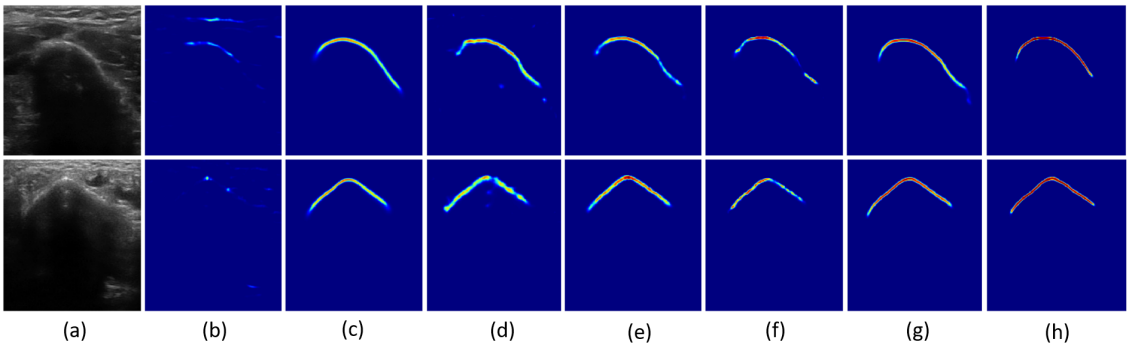

Figure 2.8: (a) Two US B-mode images of in vivo radial bones, and their corresponding segmentation outputs from the following network architectures: (b) Ronneberger et al. [16], (c) Hazirbas et al. [17], and our proposed design with (d) mid-fusion, (e) no-fusion; B-mode US images only (f) no-fusion; Local-phase filtered images only (g) early-fusion and (h) late-fusion.

instances. While the segmentation outcomes of the network in [17] and the mid-fusion network, as shown in Figure 2.9 (b) and (c) respectively, were intact, they suffered poor localization. However, the mid-fusion network had slightly better localization as the warping error indicates. On the other hand, the late-fusion network had the higher localization accuracy compared to the other networks.

2.4 Discussion and Future Work

In this study, a multimodal CNN architecture was proposed for B-mode US bone segmentation. Our network incorporated local phase images in conjunction with B-mode US data. Quantitative and qualitative validation were performed against state-of-the-art U-net [16]. It was demonstrated that incorporating local phase bone image features improves the performance of the segmentation task. Particularly, it was observed that the late fusion of spatial-phase features resulted in higher bone segmentation probability outcomes. Our future work will involve further validations prior to utilizing the proposed methods clinically. In addition, improving the computational cost of local phase feature extraction would be essential.

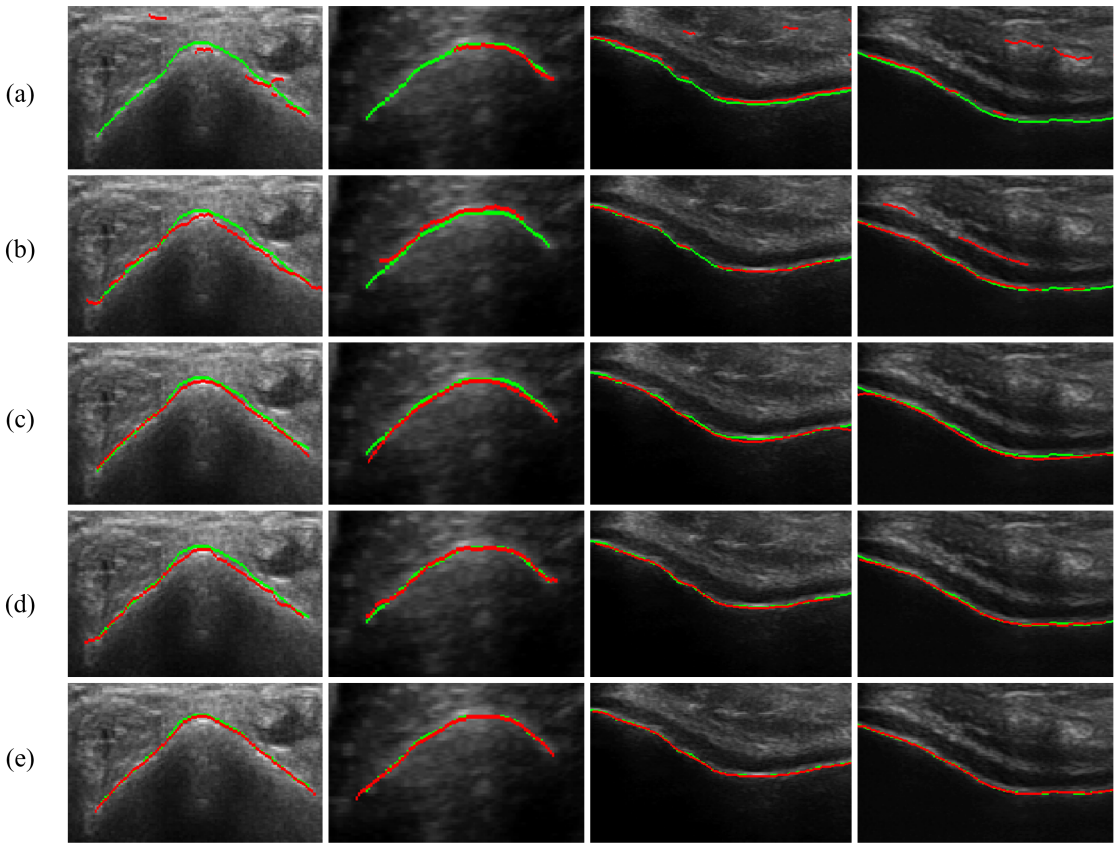

Figure 2.9: Bone localization obtained from the proposed method (red) to manual segmentation (green). The four network architectures: (a) Ronneberger et al. [16] trained with B-mode US images only, (b) Ronneberger et al. [16] trained with local-phase filtered images only, (c) Hazirbas et al. [17], (d) our mid-fusion design, and (e) our late-fusion design.

Table 2.4: Error Metrics by Bone Structure Type

Method	Dataset	IoU%	Pixel%	F-Score	Warping	Rand	AED
Ronneberger et al. [16]	Radius	45.757	69.827	0.627	0.0012175	0.40765	2.180425
Hazirbas et al. [17]	Radius	47.972	69.567	0.648	0.0001300	0.06890	0.265775
Mid Fusion	Radius	49.417	67.565	0.661	0.0002000	0.04865	0.454550
No Fusion-BM	Radius	51.378	66.349	0.678	0.0001760	0.04987	0.228760
No Fusion-LP	Radius	54.286	69.153	0.703	0.0001480	0.04379	0.218640
Early Fusion	Radius	66.369	51.954	0.795	0.0000156	0.02854	0.167855
Late Fusion	Radius	**75.452**	**37.565**	**0.860**	**0.0000100**	**0.01395**	**0.122900**
Ronneberger et al. [16]	Femur	48.367	66.020	0.652	0.0005638	0.32435	1.301475
Hazirbas et al. [17]	Femur	55.115	59.967	0.710	0.0002160	0.00045	0.163875
Mid Fusion	Femur	59.580	57.130	0.746	0.0001030	0.02385	0.229050
No Fusion-BM	Femur	60.873	59.836	0.756	0.0001020	0.01765	0.148760
No Fusion-LP	Femur	61.965	59.263	0.765	0.0001013	0.01277	0.147655
Early Fusion	Femur	69.479	59.925	0.824	0.0001014	0.00215	0.139450
Late Fusion	Femur	**84.565**	**29.562**	**0.916**	**0.0000110**	**0.00015**	**0.128175**
Ronneberger et al. [16]	Knee	51.247	62.480	0.677	0.0000218	0.01537	0.903375
Hazirbas et al. [17]	Knee	55.495	59.692	0.713	0.0000160	0.00187	0.197400
Mid Fusion	Knee	59.145	55.290	0.743	0.0000140	0.00057	0.169470
No Fusion-BM	Knee	63.767	50.657	0.778	0.0000140	0.00058	0.156390
No Fusion-LP	Knee	64.876	49.767	0.786	0.0000130	0.00049	0.155970
Early Fusion	Knee	70.168	36.981	0.822	0.0000035	0.00039	0.142789
Late Fusion	Knee	**87.695**	**27.160**	**0.934**	**0.0000001**	**0.00037**	**0.132350**
Ronneberger et al. [16]	Tibia	42.795	78.842	0.599	0.0011102	0.30112	0.393600
Hazirbas et al. [17]	Tibia	54.580	61.132	0.706	0.0009560	0.00570	0.435925
Mid Fusion	Tibia	44.895	74.072	0.619	0.0008612	0.07694	0.165350
No Fusion-BM	Tibia	57.782	65.539	0.732	0.0002879	0.06883	0.129768
No Fusion-LP	Tibia	59.767	63.824	0.748	0.0002198	0.05862	0.121876
Early Fusion	Tibia	66.148	52.764	0.794	0.0000463	0.02784	0.118921
Late Fusion	Tibia	**89.735**	**25.222**	**0.945**	**0.0000103**	**0.00240**	**0.118575**

CHAPTER 3
AUTOMATIC SEGMENTATION OF BONE SURFACES FROM ULTRASOUND USING A FILTER LAYER GUIDED CNN

In this chapter, we present our development of a block-based CNN for segmentation of bone surfaces from in vivo US scans. The novelty of our proposed design is that it also utilizes fusion of feature maps and employs multi-modal images to abate sensitivity to variations caused by imaging artifacts and low intensity bone boundaries. B-mode US images, and their corresponding local phase phase filtered images are used as multi-modal inputs for the proposed fusion network. Different fusion architectures are investigated for fusing the B-mode US image and the local phase features. The proposed methods was quantitatively and qualitatively evaluated on 546 in vivo scans by scanning 14 healthy subjects. We achieved an average F-score above 95% with an average bone surface localization error of 0.2 mm. The reported results are statistically significant compared to state-of-the-art. An improvement to the aforementioned multimodal CNN architecture in Chapter 2, using convolutional blocks instead of convolutional layers in our CNN, in this chapter we demonstrate how projection blocks can be utilized to allow semantic information to be more efficiently passed forward in the network while progressively increasing feature map sizes. In addition, we show how these computationally less expensive projection blocks can allow us to have more comprehensive feature maps. The overall result is a deeper vanishing gradient-free filter layer guided CNN architecture.

3.1 Introduction

Orthopedic procedures have been a prominent solution in treating interminable pain and disabilities, due to musculoskeletal diseases, e.g. osteoarthritis, spinal conditions, osteoporosis, and trauma injuries. In 1990, the World Health Organization reported 1.7 million

hip fractures and projected the figure to increase to 6 million by 2050 [1]. Osteoporosis and related fracture treatments costs were estimated at $\$19.1$ billion in 2004. Moreover, spine related injuries for the years 2002-2004 were estimated at $\$193.9$ billion [1, 2]. Surgical interventions may include osteotomy, fracture fixation, or placement of an implant device [3]. The need for high precision in the mentioned procedures is evidently required in order to minimize the intra- and post-operative complications. Computer assisted orthopedic surgery (CAOS) systems enable higher precision by providing surgeons, intra-operatively, real-time feedback for guidance during the procedure. Imaging is one of the most important components of any CAOS system. The standard intra-operative imaging modality in CAOS is 2D/3D fluoroscopy. Navigation during surgery is difficult using 2D fluoroscopy imaging due to limited 3D information available in 2D scans. 3D fluoroscopy systems provide a solution to this problem, however, they are twice as expensive and currently are not as widely employed as their 2D alternative. Most importantly both of these modalities operate with ionizing radiation which causes important safety concerns to both surgical team and the patient. In order to provide a safe intra-operative imaging alternative, ultrasound (US) has been incorporated into various CAOS systems. US provides real-time, non-radiation based 2D/3D imaging. However, low signal-to-noise ratio (SNR), imaging artifacts, limited field of view and being a user operated imaging modality have hindered the widespread use of US in CAOS procedures. Furthermore, the beamwidth in elevation direction strongly influences the bone surface response profile making the bone boundaries appear several mm in thickness [55]. In order to provide a solution to these difficulties, focus has been given to develop automated US bone segmentation and enhancement methods. Accurate segmentation is also very important if guidance is performed using the US data only. During procedures, such as epidural anesthesia, the guidance is achieved using the anatomical landmarks extracted from the US data.

Segmentation methods, based on image intensity and phase information, were proposed by various groups [55]. Intensity-based approaches are: (1) not robust to typical imaging

artifacts, and (2) affected by intensity variations which can be a result of changing the US machine acquisition settings or scanning patients with different body mass index (BMI). In order to provide an intensity invariant alternative, methods based on image local phase information have been proposed [55]. Although local phase methods provide more robust outcomes for enhancing the bone surfaces from the US data, compared to image intensity, it requires a post processing step for the segmentation of the bone surfaces [55, 56, 38].

Recently, machine learning methods have also been proposed for the segmentation of bone surfaces from US data. Ozdemir et al. [44] investigated the use of a supervised learning framework that combines the physical constraints of US into a graphical model for segmentation. They utilized the statistical, textural, intensity, and local phase image features. The reported bone surface localization accuracy was 0.59 mm with a computation time of 2 minutes making this method suboptimal for intra-operative use due to high computational cost. In [57], Random forest was used for classification of US bone data. The reported recall and precision rates were 0.82 and 0.84 respectively with a maximum computation time of 0.6 seconds per slice. Bone surface localization accuracy was not reported. In [41], the authors modify a deep learning network architecture, termed U-net and proposed in [16], for segmentation of bone surfaces from US data. Localization accuracy was not reported, however, the recall and precision rates for the proposed method were 0.87. In [42], a similar architecture based on U-net was investigated for segmenting vertebra bone surfaces. Reported precision and recall rates were 0.88 and 0.94, respectively. Bone surface localization accuracy was not investigated. Low quality bone surfaces were excluded from the validation and testing procedure. In [58], U-net was utilized in the development of a classification network that simultaneously perform bone segmentation with a concatenated input. The success of the deep learning methods is dependent on: (1) number of scans used for training, (2) anatomical variations (such as type of bone surfaces) present in the training data, (3) quality of the collected US data. High quality bone US data is usually defined as a high intensity bone response profile, corresponding to the bone

surface, followed by shadow region.

In this study, we propose a Convolutional Neural Network (CNN)-based approach for automated bone segmentation. Based on [45], [16], [18] and [40], we introduce a new CNN design that can perform accurate segmentation of bone structures in US images. We show that the performance of the CNN segmentation methods improves if the training is performed on data that incorporates local phase image features in addition to the intensity features of the B-mode US data. This novel approach attempts to alleviate the shortcomings of unimodal designs, and their susceptibility to noise and other imaging artifacts. Validation is performed on 546 in vivo scans obtained from 14 healthy subjects by scanning various bone surfaces. We also include quantitative and qualitative evaluation results on data sets obtained from a different US imaging platform which was not used during the training of the proposed method. Obtained results are also compared against U-net [16] trained using (1) B-mode US data only, and (2) B-mode and local phase image features.

3.2 Methods

3.2.1 Data acquisition

After obtaining the institutional review board (IRB) approval, a total of 415 B-mode US images (categorized into four groups of bone structures: radius, femur, knee, and tibia) from twelve healthy subjects, were collected using Sonix-Touch US machine (Analogic Corporation, Peabody, MA, USA). Depth settings and image resolutions varied between 3-8 cm, and 0.12-0.19 mm, respectively. In addition, a second dataset of 131 images were obtained using a handheld wireless ultrasound probe (Clarius C3, Clarius Mobile Health Corporation, BC, Canada) from two volunteers. This dataset was considered for validation/testing purposes. All the bone surfaces were manually segmented by an expert ultrasonographer.

3.3 Network architecture

Our proposed CNN architecture is based on the common contractive-expansive design, as depicted in Figures 3.1 through 3.3. The encoder maps the input image into a low-dimensional latent space, and the decoder maps the latent representation into the original space. We first resize the input B-mode US image $US(x, y)$ and its complementary local phase filtered image $LP(x, y)$ to a standardized 256×256 size. After network operations the images were resized to their original size before quantitative and qualitative validation. In our proposed design, each input image would connect to an independent primary network and a secondary network. In each network, the input image is processed through convolutional blocks, with each block consisting of several convolutional layers. Utilized in our design are four distinct blocks (colored in blue, yellow, green, and orange) that are depicted and labeled in the legend of each network design at the bottom of (Figures 3.1 through 3.3). The networks in our design incorporated skip connection and projection blocks similar to [59]. In each of the four convolutional blocks, d1 and d2 indicate the depth of each convolutional layer, while s indicates stride. Within the network design, the specific depths of each convolutional block are specified. A skip connection block consists of 1×1 convolutions before, and after a 3×3 convolution, reducing and restoring channel dimensions, respectively. Each convolution is followed by batch normalization and rectified linear unit (ReLU) activation. The output of the skip connection block is obtained by concatenating its input with the aforementioned convolutions. Our projection blocks consist of a similar structure to the skip connections, with the difference being the output is the result of concatenating the aforementioned convolutions with the projected input through a 1×1 convolution. We employ projection blocks as a means of maxpooling when a stride of 2 convolution is used. On the other hand, transposed-convolution blocks were implemented in the decoder path of each network. The design of the transposed convolution blocks are similar to the aforementioned skip and projection blocks with all convolution operations

replaced by transposed-convolutions. We also use a stride of 2 transposed convolutions to upsample the feature maps. In the primary network, the input image is a B-mode US image $US(x, y)$, while in the secondary network, the input is a local phase filtered image, $LP(x, y)$, that proceeds through the aforementioned blocks. Feature maps extracted from both networks are fused (Figures 3.1 through 3.3) in a fusion layer at various stages depending on the model. Specifically, we investigate early, mid and late fusion network models. Our early-fusion model, depicted in Figure 3.1, fuses the input B-mode US image, $US(x, y)$, and the local phase filtered image, $LP(x, y)$, at the pixel level. The fused image is then processed through a single network. In Figure 3.2, a feature level fusion model was implemented in which mid-level features from both primary and secondary networks are fused together. Finally, in Figure 3.3, a classifier level model was implemented in which high-level features from each network are concatenated. A 3×3 convolution with sigmoid activation is performed on the output of the fused layer to generate the final segmented probability distribution. Throughout our network designs, concatenation fusion is used as the fusion operation [50]. Concatenation fusion does not define any correspondence as it stacks feature maps at the same spatial locations across the feature channels. However, subsequent layers define the correspondence by learning suitable filters that weight the layers.

3.3.1 Training and testing

The performance of the proposed designs were compared against each other and the networks proposed in [16], termed U-net, and [17]. The depths of both networks in [16] and [17] were increased to a scale close to our proposed designs. In order to further validate the effectiveness of our design, we trained the U-net network proposed in [16] using: (1) Bmode-US image features only, (2) Local phase image features only, and (3) both B-mode US and local phase image features. Our proposed designs and the two networks in [16] and [17] were trained using a training set of 300 B-Mode US images (out of the 415 im-

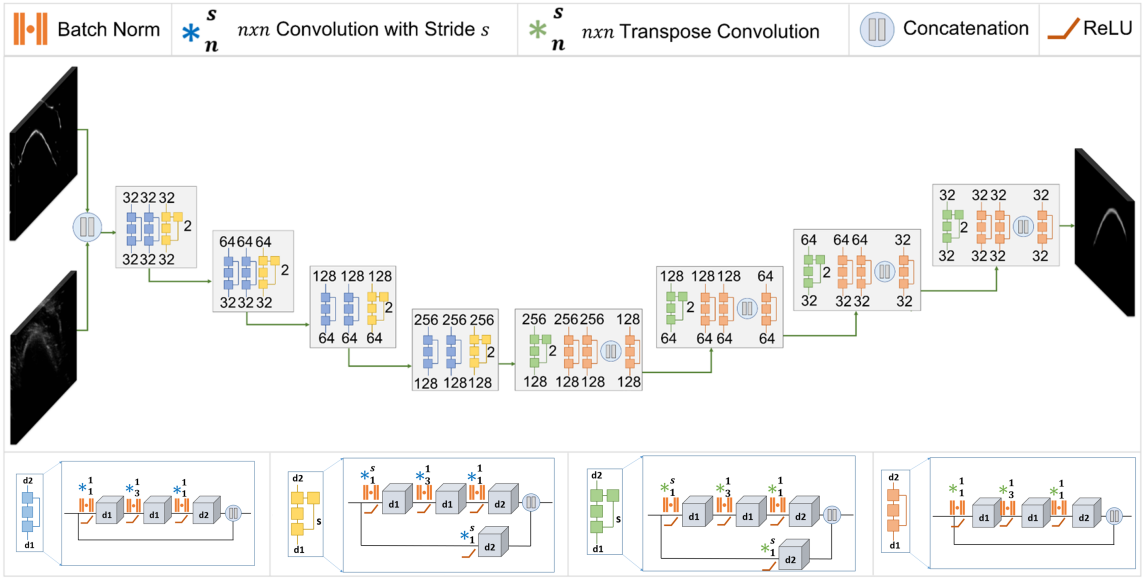

Figure 3.1: An overview of the early-fusion CNN architecture. Input B-mode US image, $US(x, y)$, is concatenated with the local phase filtered image, $LP(x, y)$, at the pixel level, and the result is processed through the network.

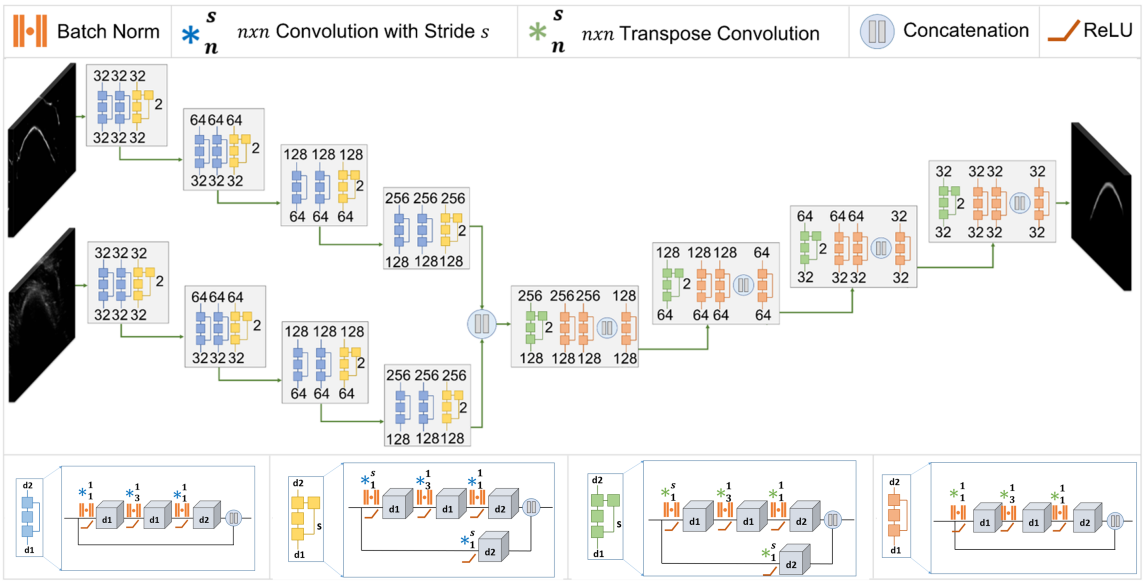

Figure 3.2: An overview of the mid-fusion CNN architecture. Input B-mode US image, $US(x, y)$, is processed through the primary encoder (bottom), while the local phase filtered image, $LP(x, y)$, is processed through the secondary encoder (top). Feature maps from both encoders are fused in a mid fusion stage, and processed through the decoder.

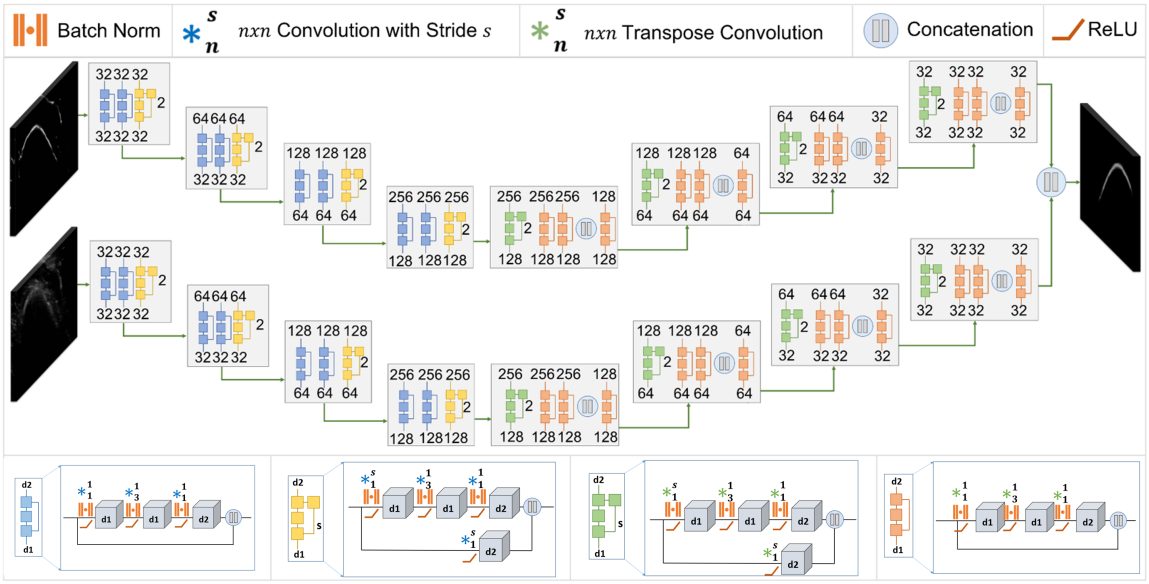

Figure 3.3: An overview of the Late-fusion CNN architecture. Input B-mode US image, $US(x, y)$, is processed through the primary network (bottom), while the local phase filtered image, $LP(x, y)$, is processed through the secondary network (top). Feature maps from both networks are fused in a late fusion stage.

ages obtained from Sonix-Touch US machine) and their corresponding local phase filtered images. The remaining 115 B-mode US images were reserved for testing the performance of the networks. During the random split of the SonixTouch dataset, same patient scans were not used for both training and testing. We repeated this process five times, with each training and testing data randomized from our datasets. In addition, we reserved another set of data for testing and validation only. This dataset consisted of 131 US scans from two different volunteers obtained using the Clarius C3 probe. The two volunteers were not part of the scanning process performed using the SonixTouch machine. Our proposed network designs were trained to minimize the cross-entropy loss. All networks were trained to 100 epochs and the best performing model was selected for each network. Based on [45, 54, 53, 52], five error metrics were calculated in our testing set; namely, F-score, Rand error, Hamming Loss, as well as the IoU and average bone surface localization error. To measure how similar any two segmentation regions in an image, the Rand error, which takes into account shifts in boundary locations, was calculated as $R_e = 1 - R_i$, where R_i is the Rand

index. Bone localization was achieved by thresholding the estimated probability segmentation map and using the center pixels along each US scanline as a single bone surface. The bone localization error is calculated as the average Euclidean distance (AED) error between the automatically segmented bone surfaces and the manual expert segmentation. The evaluation metrics were computed on the estimated probability maps, with grayscale colormaps, and compared to our manual segmentations.

3.4 Results and Discussion

Experiments were carried out using the Keras framework and Tensorflow as backend with an Intel Xeon CPU at 3.00GHz and an Nvidia Titan-X GPU with 8GB of memory. On average, our networks converge in about 6 hours during the training process. Testing was performed in real-time and on average it took 52 ms to test an image.

3.4.1 Quantitative results

The aforementioned error metrics were calculated for each of the networks, and the results are tabulated in Table 4.2. As can be seen from Table 4.2, the average numerical error calculations show that that the late-fusion design had the lowest errors, and the highest average IoU and F-scores. A paired t-test, at a $\%5$ significance level, between our designed networks and the networks proposed in [16], [17] achieved p-values less than 0.05 indicating that the improvements of our method are statistically significant. The AED results increased for all the networks analyzed when using the data obtained from the Clarius imaging platform. This is an expected results since this data was obtained from an imaging platform which was not part of the training process. However, incorporating local phase image features increases the success of the network proposed in [16].

The overall AED error for our late-fusion design is 0.1482 mm (standard deviation (SD) 0.028 mm). U-net network [16] using B-mode, local phase, and combine (B-mode and local phase features) achieved overall AED errors of 2.296 mm (SD 0.038 mm), 1.0319

Table 3.1: Error Metrics

Method	IoU%	F-Score	Rand	Hamming	AED (mm)
Dataset I - Sonix-Touch US machine					
Ronneberger [16] B-mode US (BM) only	0.864986	0.912431	0.705538	0.135013	2.4344
Ronneberger [16] Local Phase (LP) only	0.870279	0.915084	0.654647	0.129720	1.0443
Ronneberger [16] BM & LP Early Fusion	0.914163	0.944772	0.626641	0.085836	0.6375
Ronneberger [16] BM & LP Mid Fusion	0.928410	0.952721	0.655679	0.071590	0.4927
Ronneberger [16] BM & LP Late Fusion	0.948090	0.964523	0.626641	0.051913	0.2864
Hazirbas [17]	0.894572	0.931847	0.657970	0.105427	0.8582
Ours Early Fusion	0.972125	0.978072	0.448373	0.027874	0.1087
Ours Mid Fusion	0.957193	0.969739	0.484314	0.042806	0.1183
Ours Late Fusion	**0.972865**	**0.978388**	**0.439368**	**0.027134**	**0.1071**
Dataset II - Clarius C3 US probe					
Ronneberger [16] B-mode (BM) US only	0.820128	0.878605	0.647348	0.179871	2.1576
Ronneberger [16] Local-Phase (LP) only	0.869984	0.950463	0.746484	0.179871	1.0195
Ronneberger [16] BM & LP Early Fusion	0.904848	0.944324	0.760781	0.095151	0.7463
Ronneberger [16] BM & LP Mid Fusion	0.932476	0.956250	0.656760	0.067523	0.4682
Ronneberger [16] BM & LP Late Fusion	0.948085	0.964500	0.624421	0.051914	0.3755
Hazirbas [17]	0.842741	0.901746	0.670910	0.157258	0.8642
Ours Early Fusion	0.968001	0.976297	0.487246	0.031998	0.2186
Ours Mid Fusion	0.958058	0.969953	0.485360	0.041941	0.2579
Ours Late Fusion	**0.970965**	**0.977650**	**0.453752**	**0.029034**	**0.1893**

mm (SD 0.059 mm), 0.7060 mm (SD 0.05 mm) respectively. The network proposd in [17] achieved overall AED error of 0.8612 mm (SD 0.0834 mm). Again a paired t-test, at a $\%5$ significance level, between our proposed late fusion network and other network achieved p-values less than 0.05 for overall AED errors.

3.4.2 Qualitative results

Qualitative results of our early, mid, and late-fusion network designs as well as the networks in [16] and [17] are shown in Figure 3.4, where the red pixels indicate high prediction scores while blue pixels indicate low prediction scores for the segmentation. The prediction outcome when only B-mode US images were used in training, as the case in Figure 3.4

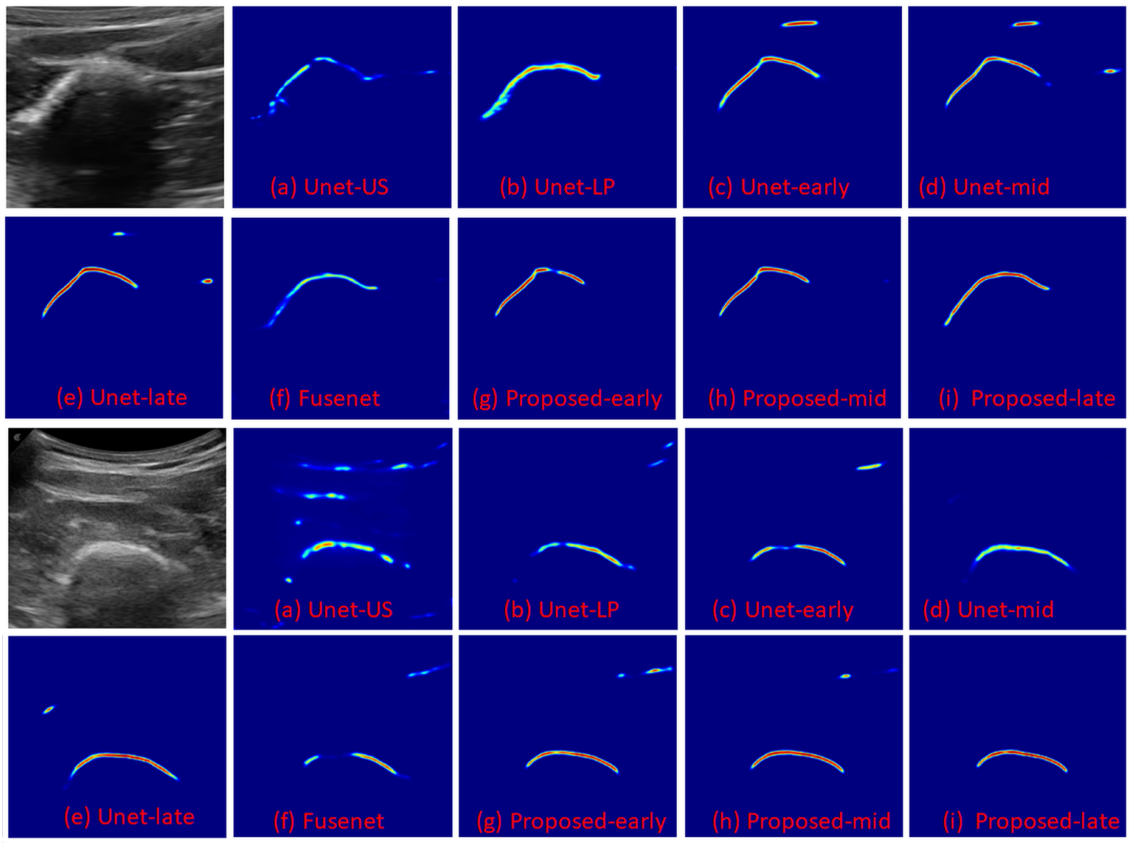

Figure 3.4: First column in vivo US B-mode images of distal radius (top), and femur (bottom). Image are obtained from the Clarius platform. Network segmentation results obtained using: Ronneberger et al. [16] trained with (a) B-mode US images only (U-net), (b) local phase filtered images only (U-LP), (c) B-mode US and local phase filtered images using early-fusion (Unet-early), (d) B-mode US and local phase filtered images using mid-fusion (Unet-mid),(e) B-mode US and local phase filtered images using late-fusion (Unet-late). (f) Hazirbas et al. [17] trained with both B-mode US and Local-phase filtered images (Fusenet). Our proposed designs (g) early-fusion, (h) mid-fusion, and (i) late-fusion.

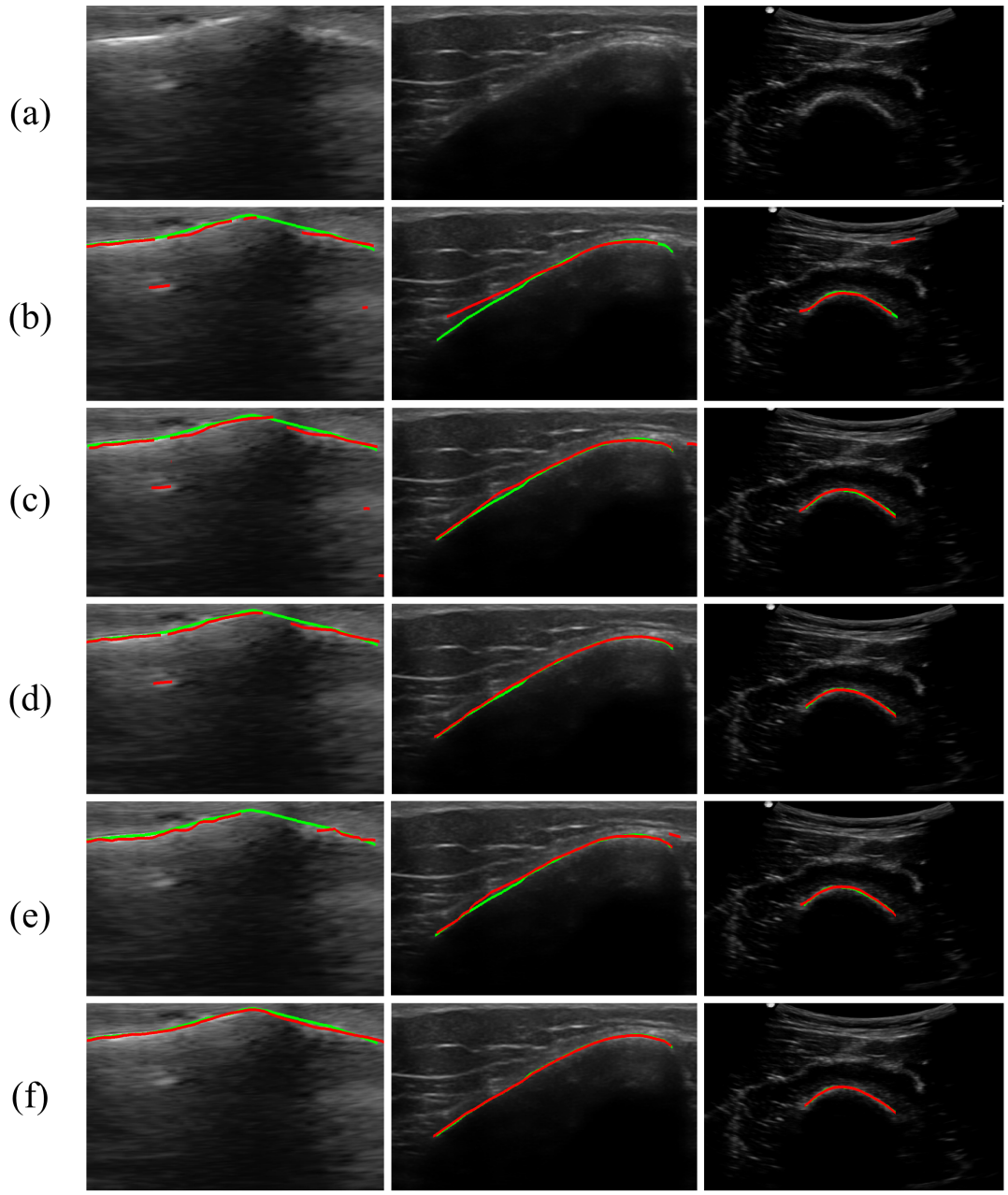

Figure 3.5: Bone localization obtained from the proposed method (red) to manual expert localization (green). (a) In vivo B-mode US image of distal radius (top) and femur (bottom). (b) Ronneberger et al. [16] trained with B-mode US images only, (c) Ronneberger et al. [16] trained with local phase filtered images only, (d) Ronneberger et al. [16] trained with B-mode US and local phase filtered images,(e) Hazirbas et al. [17], and (f) our late-fusion design.

(b) for [16], had the lowest probability distribution amongst all others. The inadequate segmentation performance may be attributed to the nature of the US images used in the testing process. For low quality US scans, where the bone surface has a low intensity profile and high intensity soft tissue interfaces appearing above the bone surface, the performance of the network proposed in [16] declines. The importance of collecting high quality US data and its affect on the segmentation outcome was also discussed previously in [42] who proposed a similar network architecture for segmenting vertebra bone surfaces from US data. Figure 3.4 (d) shows an improved prediction outcome when both B-mode US and Local-phase filtered images are used to train the U-net architecture proposed in [16]. The network in [17] had a lower probability distribution than our proposed fusion networks. This is because in the network proposed in [17], the fusion performed at the feature level is considered a slow fusion in which multiple feature maps are fused throughout the encoder. As shown in Figure 3.4 (f), early fusion outperforms the network in [17] since the fusion happens at the pixel level in which the fused image would possess enhanced bone surfaces while the soft tissue interfaces remain unaltered. Investigating Figure 3.4 (first and third rows) we can also see that all the networks perform better when the testing data is from the same imaging platform where the training data is obtained. However, when using test data obtained from an imaging platform which the networks have not seen during training the performance decreases. However, we can still infer that our network designs, compared to U-net [16] and [17], perform better resulting with segmentation outcomes with high probability.

Bone localization results, against expert manual localization, are presented in Figure 3.5. The specific B-mode US data presented in this figure (Fig. 3.5-a) show low quality bone scans. Investigating the localization results we can infer that U-net [16] trained only using B-mode data achieves the worst performance: large gap from the expert localization, missing bone boundaries, false positive bone localizations. Although the performance increases when the same network is trained together with B-mode and local phase features

(Fig. 3.5-d) false and true positive localization is still visible. Qualitatively our late-fusion design achieves the best performance for this dataset.

3.5 Discussion and Conclusions

In this study, three CNN architectures, for the task of bone segmentation from US data, were proposed. Our networks incorporate local phase images in conjunction with B-mode US data. We have investigated how to combine information from local phase images and B-mode US data by analyzing different fusion strategies. Our results demonstrate that for the task of bone segmentation fusing B-mode US and local phase features at a later stage outperforms early and mid fusion, specifically for the dataset obtained from Clarius C3 US probe. Since local phase image features enhance the bone surface response in the US data, the B-mode US data and local phase image features are less correlated in the low level features. The proposed late level fusion network models the correlations and interactions between high level features of each modality, outperforming the other fusion networks. A similar investigation can also be observed with the U-net network late fusion design [16] (Table.1). We also show that incorporating local phase bone image features, using three different stages of fusion, improves the performance of state-of-the-art U-net network [16]. Conducted quantitative studies show significant improvement of our network with late fusion over state-of-the-art CNN methods [16].

In our network architecture we use convolutional/projection blocks. Our projection blocks allow semantic information to be more efficiently passed forward in the network while progressively increasing feature map sizes, compared to simple convolutions which is used in the U-net design [16]. The projection blocks allow us to have more comprehensive feature maps. This is one of the reasons why our fusion networks outperform fusion networks of U-net design [16].

One of the drawbacks of the proposed work is the computational time required for the extraction of local phase image features. This takes on average 1 second (MATLAB

implementation) which needs to be improved for real-time CAOS procedures where US is used as an intra-operative imaging modality. Furthermore, during this work the expert manual segmentation was performed by a single expert user. The effect of intra- and inter-user expert bone segmentation on the segmentation results is also crucial. Our future work will involve (1) extensive clinical validation of the proposed method, (2) improving the computational cost of local phase feature extraction, (3) inter- and intra-user variability analysis for expert bone segmentation, and (4) extension of our network architecture to process volumetric US data [33].

CHAPTER 4
BONE SHADOW SEGMENTATION FROM ULTRASOUND DATA FOR ORTHOPEDIC SURGERY USING GAN

In this chapter, we present a computational method, based on a novel generative adversarial network (GAN) architecture, to segment bone shadow images from in vivo US scans in real-time. We also show how these segmented shadow images can be incorporated, as a proxy, to a multi-feature guided convolutional neural network (CNN) architecture for real-time and accurate bone surface segmentation. Quantitative and qualitative evaluation studies are performed on 1235 scans collected from 27 subjects using two different US machines. Finally we provide qualitative and quantitative comparison results against state-of-the-art GANs. We have obtained mean dice coefficient ($\pm$ standard deviation) of % 93 ($\pm$ 0.02) for bone shadow segmentation, showing that the method is in close range with manual expert annotation. Statistical significant improvements against state-of-the-art GAN methods (paired t-test p ¡ 0:05) is also obtained. Using the segmented bone shadow features average bone localization accuracy of 0.11mm ($\pm$ 0.16) was achieved. Reported accurate and robust results make the proposed method promising for various orthopedic procedures. Although we did not investigate in this work, the segmented bone shadow images could also be used as an additional feature to improve accuracy of US-based registration methods. Further extensive validations are required in order to fully understand the clinical utility of the proposed method.

4.1 Introduction

Imaging has been an integral component of various surgical and non-surgical orthopedic procedures such as total knee replacement (TKR), intramedullary nail locking for femoral shaft fractures, pedicle screw insertion for spinal fusion surgery, lumbar neuraxial anesthe-

sia, and epidural analgesia [3]. Current practice during these procedures relies on intra-procedure 2D fluoroscopy as the main imaging modality for localization and visualization of bones, fractures, implants, and surgical tool positions. However, with such projection imaging, surgeons and clinicians typically face considerable difficulties in accurately localizing bone fragments in 3D space and assessing the adequacy and accuracy of the procedure. This problem has been overcome with 3D fluoroscopy units, however, they are twice as expensive and not widely available as standard 2D units. Finally, fluoroscopy involves significant radiation exposure [3]. The limits to exposure to ionizing radiation should be kept at minimal in order to avoid potential long-term complications. In order to overcome some of these limitations and provide a safe alternative, 2D/3D US has emerged as a safe alternative while remaining relatively cheap and widely available [55]. US image data, however, is typically characterized by high levels of speckle noise, reverberation, anisotropy and signal dropout which introduce significant difficulties during interpretation of captured data. Limited field-of-view and being a user dependent imaging modality causes additional difficulties during data collection since a single-degree deviation angle by the operator can reduce the signal strength by 50% [55]. In order to overcome these difficulties automatic bone segmentation [55] and registration [60] methods have been developed. Most recently, methods based on deep learning have achieved successful results for segmenting bone surfaces [61, 42, 43, 58]. However, these methods require large amounts of training data and accuracy decreases if the quality of the testing data is low or if testing data comes from a different vendor machine. In the context of bone imaging using US, high quality data represents high intensity bone surface followed by a low intensity region referred to as shadow region. Difficulties in acquiring high quality US images is an ongoing limitation of current US guided orthopedic procedures.

Acoustic shadows occur at the interfaces where there is a high impedance difference such as air-tissue, tissue-bone, and tissue-lesion. Bone shadow information can aid in the interpretation of the collected data and has been incorporated as an additional feature to

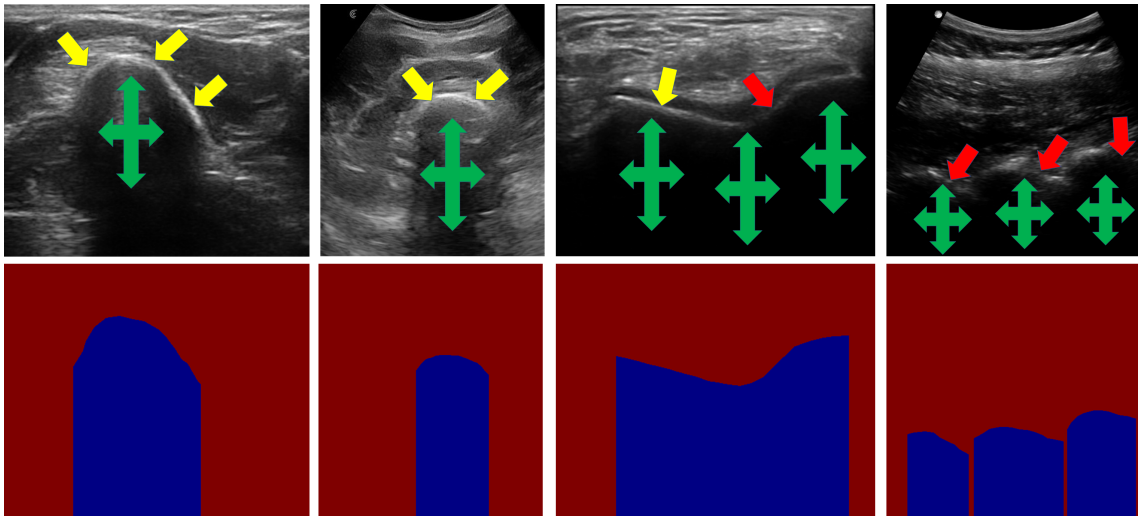

Figure 4.1: Top row: From left to right in vivo B-mode US image of distal radius, femur, knee, and spine respectively. Yellow arrows point to high intensity bone features. Red arrows point to the problematic low intensity bone features due to misalignment of the transducer or complex shape of the anatomy. Green arrow quads show the shadow region. Bottom row:Manually segmented gold standard shadow images corresponding to B-mode data shown in the top row. In all the images blue color coded region is the shadow region and red color coded region is the soft tissue interface.

improve the segmentation of bone surfaces from US data [62, 63, 55, 58]. Real-time feedback of bone shadow information can also be used to guide the clinician to a standardized diagnostic viewing plane with minimal artifacts. Finally, shadow information can also be used as an additional feature for registering CT, MRI or statistical shape models (SSM) to US data [60]. However, poor transducer contact or wrong orientation of the transducer with respect to the imaged anatomy can lead to poor shadow appearance and resulting in misinterpretation of anatomy and failure of the computational method using the shadow feature (Fig. 4.1). Therefore, the enhancement of shadow regions has been investigated and practical solutions have been offered.

Several groups have proposed computational methods to improve the appearance of shadow regions from US data. Karamalis et. al. [37] have proposed a random-walk geometric technique, based on image intensity, that models the propagation path of an US signal along the scanline. The generated images were termed confidence map (CM) im-

ages. Shadow regions were extracted from the CM images by intensity thresholding. This approach was later extended for processing radio-frequency US data [64]. In [65], shadow images of the brain were extracted by entropy analysis along the scanline. Pixels with low entropy would be selected to form the shadow image[65]. The method was later incorporated into a spinous process segmentation framework [62]. In [66], statistics of B-mode and radio frequency (RF) US data were investigated and used for shadow detection. Mean dice similarity coefficient (DSC) of 0.90 and 0.87 were obtained for the RF and B-mode algorithms. Processing time was not reported. Although promising results in these earlier works were achieved, intensity-based approaches are not robust to typical imaging artifacts and affected by intensity variations. Changing the US machine acquisition settings, sub-optimal orientation of the transducer concerning the imaged anatomy, imaging complex shape anatomy (such as spine), or scanning patients with different body mass index results in the collection of low quality US data (Fig. 4.1) and decrease the success of intensity-based approaches. RF-based shadow detection overcomes some of the difficulties of intensity approaches, however, they require special hardware, or software, to access RF signal domain which is not available in most clinical US machines. In order to provide an intensity invariant alternative, methods based on local phase image information have been proposed for the enhancement of bone shadow region [55]. The method proposed in [55] uses local phase image features as an input to a L1 norm-based contextual regularization method which emphasizes uncertainty in the shadow regions. Quantitative analysis, performed on a manually selected region of interest (ROI) achieved a mean DSC of 0.88. The mean computation time was 9.3 seconds making the method not suitable for real-time applications. In [67] a weakly supervised method for acoustic confidence estimation for shadow regions from fetal US data was proposed. In particular, a shadow-seg module to extract generalized shadow features for a large range of shadow types in fetal US images under limited weak manual annotations was presented. Both a classification and a segmentation networks with attention layer mechanism were used. The reported average DSC,

Recall, and Precision were 0.71, 0.72, 0.73 respectively.

In this paper, we propose a conditional GAN(cGAN)-based method for accurate real-time segmentation of bone shadow regions from in vivo US scans. Our specific contributions include: (1) A novel GAN architecture designed to perform accurate, robust and real-time segmentation of bone shadow images from in vivo US data. (2) We show how the segmented bone shadow regions can be used as an additional proxy to improve bone surface segmentation results of a multi-feature guided (CNN) architecture [61]. The significance of using shadow features-based segmentation is that they can be generated in real time as opposed to local phase image-based methods [61] which takes around one second. (3) We evaluate the proposed method on extensive in vivo data obtained from 27 volunteers using two different US imaging systems. We provide quantitative evaluation results against state-of-the-art GAN architectures.

4.2 Methods

4.2.1 Data acquisition

Upon obtaining the approval of the institutional review board (IRB), two imaging devices were used to collect data from 27 healthy subjects. Depth settings and image resolutions varied between 3-8 cm, and 0.12-0.19 mm, respectively:

1. Sonix-Touch US machine (Analogic Corporation, Peabody, MA, USA) with a 2D C5-2/60 curvilinear probe and L14-5 linear probe. Using this device we have collected 1000 scans from 23 subjects.

2. Clarius C3 hand-held wireless ultrasound probe (Clarius Mobile Health Corporation, BC, Canada). Using this device we have collected 235 scans from 4 subjects.

All the collected scans were scaled to a standardized size of 256×256. The bone surfaces were manually segmented by an expert ultrasonographer. Gold standard bone shadow images were constructed automatically by investigating the intensity values from the manually

segmented bone surfaces in the scanline direction. Region below the manually segmented bone surface is identified as shadow region. In total, we had 1235 B-mode US images categorized into four groups of bone structures: radius, femur, spine and tibia. We have performed 5-fold cross validation on the Sonix-Touch data. No scans from one patient appeared in more than one-fold. Training of the network architectures was performed using the Sonix-Touch data only. All the data, 235 scans in total, obtained from the Clarius C3 probe were used as test data.

4.2.2 Network architecture

Our architecture is based on the common GAN layout consisting of two co-existing neural networks; a generator G that attempts to generate synthetic samples and a discriminator D that tries to discriminate between generated synthetic samples and real ones [68]. In this work, we adopt the conditional aspect presented by [69] with our generator G and discriminator D both incorporating additional information into account. Our proposed cGAN-based bone shadow segmentation and bone surface segmentation network architecture is shown in Figure 4.2. The training of our proposed cGAN architecture follows the typical optimization problem [69] such that the discriminator D is trying to maximize and the generator G is trying to minimize the following objective $\mathcal{G}$:

$$\mathcal{G} = \arg \min_{G} \max_{D} \mathbb{E}_{BM,GS} \left[\log D(BM, GS)\right] + \mathbb{E}_{BM,z} \left[\log \left(1 - D(BM, G\left(BM, z\right))\right)\right] + \lambda \mathbb{E}_{BM,GS,z} \left[\|GS - G\left(BM, z\right)\|_1\right]$$

in which GS represents gold standard shadow images, z represents Gaussian noise in initial training but was applied as dropout on some layers in the convolution blocks, BS represents the segmented bone shadow image, GS represents the gold standard bone shadow image, and BM represents the B-mode US image. Different from a traditional GAN architecture, our actual generator G model was conditioned on the in vivo B-mode US image,

BM, and is additionally tasked to generate BS images that are as close as possible to the GS images with the introduction of the L1-distance term as shown in the equation above. Our generator architecture is based on the common contractive-expansive design where the encoder maps the input image into a low-dimensional latent space, and the decoder maps the latent representation into the original space. It is trained to generate bone shadow, BS, images. However, unlike [69] where the generator was based on [16], we employ a different structure for the generator. Similar to [61], the input is processed through convolutional blocks, with each block consisting of several convolutional layers. We incorporated skip connection and projection blocks similar to [59]. Our skip connection blocks, denoted as S, consist of a 1×1 convolution, a 3×3 convolution, and another 1×1 convolution with each convolution operation followed by batch normalization and leaky rectified linear unit (Leaky ReLU) activation. This process reduces and restores channel dimensions. In our design Leaky ReLU was used in the encoder and decoder. In [70], it has been shown that the Leaky ReLU achieves lower training and test errors compared to ReLU. Furthermore, Leaky ReLU attempts to overcome the 'dying ReLU (vanishing gradient)' problem by maintaining a small slope in the negative portion while training the piecewise constant gradient, making the network converge faster during training. This informed our choice of Leaky ReLU. A concatenated input and the aforementioned convolutions produce the output. As for our projection blocks, denoted as P, we add a 1×1 convolution to the projected input, and the rest is similar to the skip connection blocks. In the decoder, we replace all convolution operations by transposed-convolutions. We also use a stride of 2 transposed convolutions to upsample the feature maps. Therefore, these skip and projection blocks with transposed-convolutions are denoted as $S'$and P' respectively. Additionally, one difference in the decoder is that the batch normalization is followed by a dropout layer with a dropout rate of 50%. The architecture of the generator can be summarized as:

- **encoder:** $S32$ $S32$ $P32$ - $S64$ $S64$ $P64$ - $S128$ $S128$ $P128$ - $S256$ $S256$ $P256$ - $S512$ $S512$ $P512$

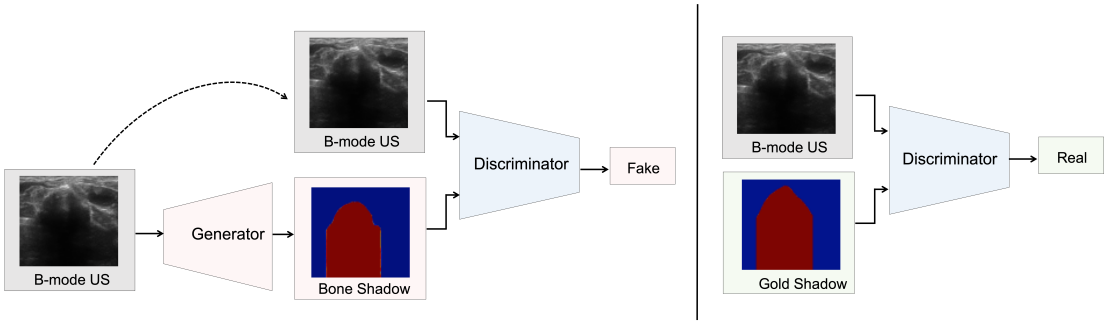

Figure 4.2: Our proposed conditional GAN in which the discriminator learns to classify between real (gold standard shadow images, GS) and fake (generated bone shadow images, BS).

- **decoder:** $S'512$ $S'512$ $P'512$ - $S'256$ $S'256$ $P'256$ - $S'128$ $S'128$ $P'128$ - $S'64$ $S'64$ $P'64$ - $S'32$ $S'32$ $P'32$

In our discriminator model a two-input $N \times N$ PatchGAN-like discriminator [69] was used to essentially classify $N \times N$ patches of the input image as real or synthetic. Like the aforementioned generator, our discriminator architecture consists of five convolutional blocks, where a final convolution is applied to the last layer to map the 1-dimensional output before applying a Sigmoid function. Each batch normalization was followed by 0.2-slope Leaky ReLU. An Adam solver with a 0.0002 learning rate was used and the structure of the discriminator can be expressed as follows:

- **discriminator:** $S32$ $S32$ $P32$ - $S64$ $S64$ $P64$ - $S128$ $S128$ $P128$ - $S256$ $S256$ $P256$ - $S512$ $S512$ $P512$

While our proposed cGAN architecture was used to segment bone shadow regions BS, our bone surface segmentation network with its dual input proposed in [61] was used in our model to localize bone structures. The B-mode US image BM, and the segmented bone shadow image BS were used as input to our multi-feature CNN architecture. Feature maps extracted from both images are fused in a fusion layer at early (pixel level), mid (feature level) and late (classifier level) stages. Concatenation fusion was used as the fusion

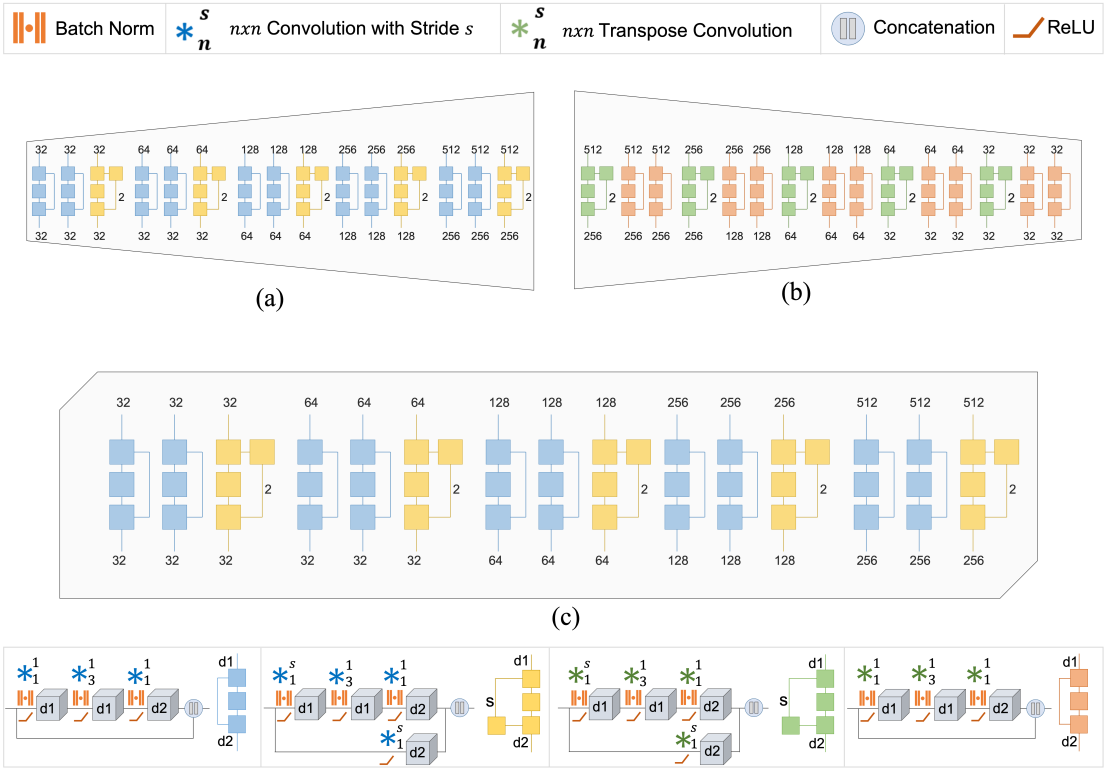

Figure 4.3: An overview of our proposed cGAN architecture with its (a) generator's encoder consisting of ten skip connection blocks (blue), in addition to five projection blocks (yellow) and (b) generator's decoder consisting of ten transposed skip connection blocks (orange), in addition to five transposed projection blocks (green). Depths of each convolutional layer are indicated in each block by d_1 and d_2. (c) Our proposed patchGAN-like discriminator.

operation [61], which does not define any correspondence as it stacks feature maps at the same spatial locations across the feature channels.The multi-feature CNN architecture was trained separately from our proposed cGAN architecture using cross-entropy loss.

4.2.3 Quantitative evaluation

Bone shadow segmentation: The performance of our proposed design was compared against state-of-the-art GAN networks proposed in [71] and [69]. The depths of the networks were increased to a scale close to our proposed design. The bone shadow regions were also segmented, from the test data, using the local phase image-based bone shadow

enhancement method proposed in [45]. In order to show the effectiveness of discriminator network we have obtained bone shadow segmentation results by only training our proposed generator network. Finally, to show the improvements achieved using a cGAN architecture over a traditional CNN architecture we trained the U-net network, proposed in [16], using B-mode US image features and gold standard bone shadow images. Based on [45, 54, 53, 52], four error metrics were calculated in our testing set: Dice, Rand error, Hamming Loss, as well as the intersection over union (IoU). The evaluation metrics are computed on the estimated probability maps, with grayscale color maps, and compared to the gold standard bone shadow images.

Bone surface segmentation: Bone shadow images segmented using our proposed design, [71] and [69], were used as an additional feature to our multi-feature CNN architecture [61] for bone surface segmentation. Our method utilizes fusion of feature maps obtained from B-mode US data and bone shadow images. During the evaluation studies we investigate different fusion architectures: early, mid and late fusion [61]. We also investigate bone surface segmentation results if gold standard bone shadow images are used as an additional feature. The bone segmentation networks were trained to minimize the cross-entropy loss. We have used Adam Optimizer with batch size of 8 and a learning rate of 0.0002 for 36,000 iterations. In addition to the previously error metrics explained in this section, we also evaluate the average Euclidean distance (AED) error for the task of bone segmentation. AED was calculated between the automatically segmented bone surfaces and the manual expert segmentation [61].

4.3 Results

Our experiments were conducted using the Keras framework and Tensorflow as backend with an Intel Xeon CPU at 3.00GHz and an Nvidia Titan-X GPU with 8GB of memory. Our network converged in about 8 hours during the training process. Testing on average

took 54 milliseconds in total for bone shadow and bone surface segmentation.

4.3.1 Quantitative results

Bone shadow segmentation: Table 4.1 shows the performance difference of bone shadow segmentation methods investigated. Overall our method outperforms previous state-of-the-art GAN architectures and the local phase-based bone shadow enhancement method [45]. The local phase image based method proposed in [45] achieved the lowest DSC value (0.28). However, we would like to mention that in the original work of [45] a ROI, covering a bone interface spanning the full width of the image, was selected during quantitative evaluation. During our analysis we did not select a ROI and rather used the full B-mode US image. Our generator network, without the discriminator, achieved average dice value or 0.67. While adding the discriminator resulted in 39% improvement in Dice value. Our proposed cGAN architecture achieves 8% and 3% improvement, in DSC value, over the state-of-the-art GAN architectures proposed in [71], [69] respectively. A paired t-test, for IoU, DSC and AED results at a $\%5$ significance level, between our proposed network and the networks in [71], [69] achieved p-values less than 0.05 indicating that the improvements of our method are statistically significant. The improvement over the U-net architecture [16] was 46% for DSC value. We have also observed that our generator network, without discriminator, outperforms U-net [16] by 6% in Dice value.

Bone surface segmentation: Quantitative results for bone surface segmentation are presented in Table 4.2. The average numerical error calculations show that that the late-fusion design had the lowest errors, and the highest average IoU and Dice (Table 4.2). A paired t-test, for IoU, DSC and AED results at a $\%5$ significance level, between our proposed network and the networks in [71], [69] achieved p-values less than 0.05 indicating that the improvements of our method are statistically significant. There was no statistical significance when using gold standard bone shadow images and the bone shadow images generated using the proposed design for late fusion design. When using local phase image features as

an additional feature for our multi-feature CNN architecture [61] the AED error was 0.30 mm compared to 0.11 mm when using the generated bone shadow images.

Table 4.1: Bone Shadow Segmentation Error Metrics

Method	IoU%	Dice	Rand	Hamming
Dataset I - Sonix-Touch US machine				
LP-based transmission maps [45]	0.2350	0.2186	0.9993	0.8722
Ronneberger et. al. [16]	0.5242	0.6504	0.9897	0.5649
Generator network only	0.5927	0.6839	0.9867	0.4019
Radford et. al. [71]	0.8015	0.8972	0.6371	0.1951
Isola et. al. [69]	0.8628	0.9404	0.6364	0.1501
Ours	**0.9277**	**0.9603**	**0.4841**	**0.0873**
Dataset II - Clarius C3 US probe				
LP-based transmission maps [45]	0.2670	0.2802	0.9983	0.8722
Ronneberger et. al. [16]	0.4726	0.6374	0.9857	0.5272
Generator network only	0.5157	0.6755	0.9831	0.4848
Radford et. al. [71]	0.7965	0.8620	0.6685	0.2034
Isola et. al. [69]	0.8424	0.9015	0.6730	0.1575
Ours	**0.9023**	**0.9354**	**0.5990**	**0.0976**

4.3.2 Qualitative results

Qualitative results of our proposed model are shown in Figure 4.4. We demonstrate five examples of in vivo US B-mode images bone types, namely: femur, tibia, radius, knee, and spine, where red pixels indicate high prediction scores while blue pixels indicate low prediction scores for the prediction. Gold standard bone shadow images obtained by an expert are displayed followed by generated bone shadow results obtained using the convolutional network presented by Ronneberger et al. [16] and generative networks in [71], [69] and our proposed model, as shown in Figure 4.4 (d) through (g). In Figure 4.4-(c), we demonstrate shadow results obtained using local phase-based ultrasound transmission

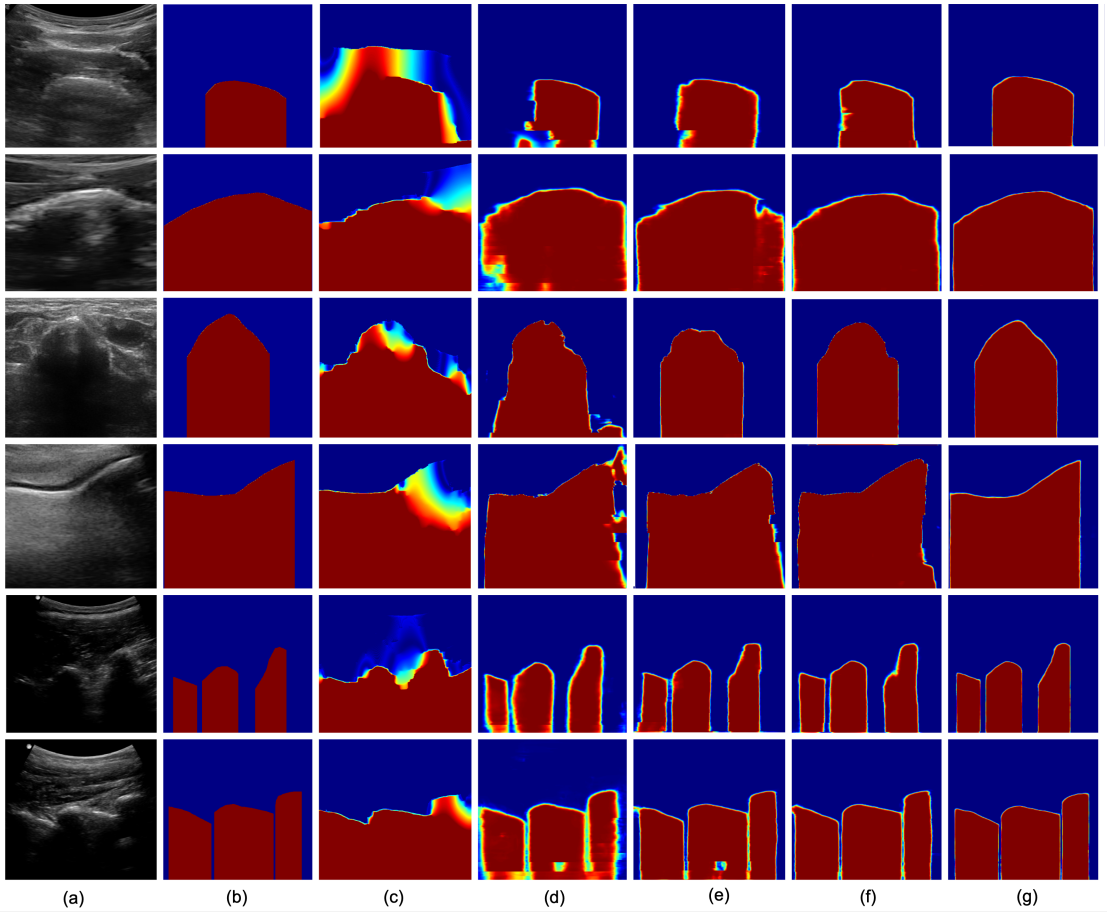

Figure 4.4: Qualitative results for bone shadow segmentation. (a) In vivo B-mode US images of femur, tibia, radius, knee, and spine. (b) Gold standard bone shadow images. (c) Bone shadow results obtained using local phase-based ultrasound transmission maps method presented in [45]. (d) Bone shadow results obtained using Ronneberger et al. [16] (e) Bone shadow results obtained using Radford et al. [71] (f) Bone shadow results obtained using Isola et al. [69]. (g) Bone shadow results obtained using our proposed cGAN.

maps method presented in [45]. Investigating the qualitative results we can conclude that our proposed method segments bone shadow images with minimal artifacts.

In Figure 4.5, early, mid, and late-fusion bone surface localization results are shown. During this qualitative evaluation we have used bone shadow images generated from our proposed cGAN architecture, together with the B-mode US data, as an input to fusion networks designed using U-net architecture [16] and our previously proposed architecture [61]. The results in Figure 4.5-(a) through (c) presented inadequate segmentation performance which may be attributed to the nature of the US images used in the testing process. For low quality US scans, where the bone surface has a low intensity profile and high intensity soft tissue interfaces appearing above the bone surface, the performance of the network proposed in [16] declines. This demonstrates the advantage of using skip and projections convolutional blocks instead of convolutional layers. Overall late fusion operation outperforms early and mid level fusion.

4.4 Discussion and Conclusions

A method, based on a novel GAN, for real-time and accurate segmentation of bone shadow regions from in vivo US scans was proposed. Our model has two main networks: (1) a cGAN to generate bone shadow images and (2) a segmentation network that will take the generated bone shadow data in conjunction with B-mode US data for localization of bone surfaces. Our integral component of building the generator and discriminator was the skip and projection blocks. To the best of our knowledge, this was not previously investigated in the community. We also would like to mention that this is the first work proposing a novel cGAN architecture for the task of bone shadow segmentation. The projection blocks allow semantic information to be more efficiently passed forward in the network while progressively increasing feature map sizes, compared to simple convolutions which is used in many designs including in [16]. By implementing these projection blocks, we allow to have more comprehensive feature maps that improve the bone shadow generation. We

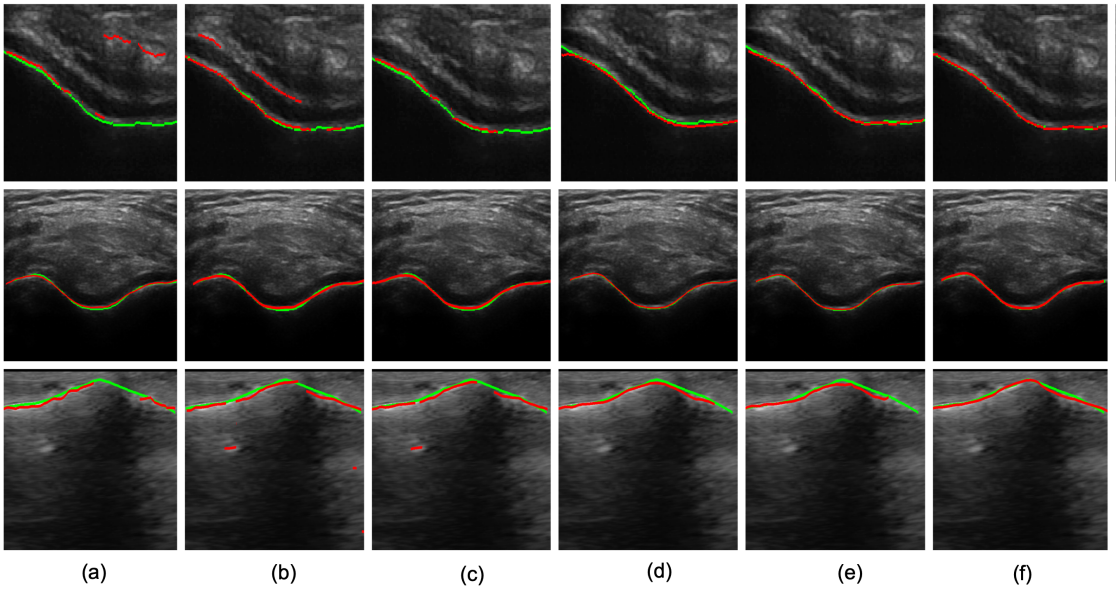

Figure 4.5: Bone segmentation results. Bone surfaces segmented using automated methods are shown as red color coding while manual expert segmentation surfaces are shown in green color coding. B-mode in vivo images and their bone shadow images counterparts (generated using our proposed model) were fused at an early, mid, and late stage [61]. (a) Ronneberger et al. [16] (early), (b) Ronneberger et al. [16] (mid), (c) Ronneberger et al. [16] (late), (d) our design in [61] (early), (e) our design in [61] (mid), and (f) our design in [61] (late)

have also extended the depth of the discriminator used in the state-of-the-art [69]. This is one of the reasons why our cGAN outperformed other state-of-the-art networks on this testing data set. Based on these results, we can conclude that having a cGAN with prior information can significantly improves the results for the task at hand. In this study we have also shown the importance of adverserial training. The success of well trained CNN architectures is effected if the architecture is deployed on test data coming from different centers, vendors, or changing acquisition parameters. For US data, even when the machine is from the same vendor the image acquisition settings can be adjusted from one scanning procedure to the next. BMI of the patient, orientation of the transducer with respect to the imaged anatomy will also change the appearance of the collected data drastically. We have shown that GAN are more robust to these conditions. We have also investigated how to combine information from bone shadow and B-mode US data by analyzing different fusion strategies. Our results demonstrate that for the task of bone segmentation fusing B-mode US and bone shadow features at a later stage outperforms early and mid fusion, specifically for the dataset obtained from Clarius C3 US probe. One of the advantages of the proposed work is that bone shadow features are obtained instantaneously making the computational time required suitable for real-time applications. Our future work will involve (1) extensive clinical validation of the proposed GAN-based method on data obtained from subjects who have differing pathology in their bone such as fracture or bone deformity such as scoliosis. We will also extend our network architecture to process volumetric US data [33].

Table 4.2: Bone Segmentation Error Metrics. BM: B-mode US image, BS: bone shadow image, LP: local phase image, GS: gold standard image

Method	IoU%	Dice	Rand	Hamming	AED (mm)
Dataset I - Sonix-Touch US machine					
Radford et. al. [71] BM & BS Early Fusion	0.8511	0.9368	0.7507	0.1003	0.5863
Radford et. al. [71] BM & BS Mid Fusion	0.8735	0.9166	0.6201	0.1297	0.3972
Radford et. al. [71] BM & BS Late Fusion	0.8834	0.9223	0.5961	0.1165	0.3258
Isola et. al. [69] BM & BS Early Fusion	0.8877	0.9248	0.5847	0.1122	0.3105
Isola et. al. [69] BM & BS Mid Fusion	0.8958	0.9294	0.5646	0.1041	0.2976
Isola et. al. [69] BM & BS Late Fusion	0.9076	0.9368	0.5430	0.0923	0.1776
Ours BM & LP Late Fusion	0.8892	0.9354	0.8290	0.1107	0.3059
Ours BM & GS Late Fusion	0.9779	0.9826	0.3062	0.0220	0.1059
Ours BM & BS Early Fusion	0.9670	0.9746	0.5775	0.0329	0.1164
Ours BM & BS Mid Fusion	0.9694	0.9770	0.4668	0.0305	0.1089
Ours BM & BS Late Fusion	**0.9803**	**0.9833**	**0.3644**	**0.0196**	**0.1032**
Dataset II - Clarius C3 US probe					
Radford et. al. [71] BM & BS Early Fusion	0.8388	0.8995	0.6782	0.1611	0.8542
Radford et. al. [71] BM & BS Mid Fusion	0.8513	0.9369	0.7512	0.1001	0.6753
Radford et. al. [71] BM & BS Late Fusion	0.8731	0.9356	0.7474	0.1029	0.4227
Isola et. al. [69] BM & BS Early Fusion	0.8655	0.9361	0.7496	0.1019	0.3814
Isola et. al. [69] BM & BS Mid Fusion	0.8986	0.9409	0.8105	0.1013	0.2669
Isola et. al. [69] BM & BS Late Fusion	0.9146	0.9442	0.5543	0.0853	0.1973
Ours BM & LP Late Fusion	0.8730	0.9250	0.8390	0.1269	0.4215
Ours BM & GS Late Fusion	**0.9695**	**0.9781**	**0.4353**	**0.0304**	**0.1106**
Ours BM & BS Early Fusion	0.9315	0.9555	0.5347	0.0684	0.1655
Ours BM & BS Mid Fusion	0.9526	0.9692	0.5081	0.0473	0.1306
Ours BM & BS Late Fusion	0.9625	0.9752	0.4908	0.0374	0.1129

CHAPTER 5
GAN-BASED REALISTIC BONE ULTRASOUND IMAGE AND LABEL SYNTHESIS FOR IMPROVED SEGMENTATION

In this chapter, we propose a computational method, based on a novel generative adversarial network (GAN) architecture, to (1) produce synthetic B-mode US images and (2) their corresponding segmented bone surface masks in real-time. We show how a duality concept can be implemented for such tasks. Armed by two convolutional blocks, referred to as self-projection and self-attention blocks, our proposed GAN model synthesizes realistic B-mode bone US image and segmented bone masks. Quantitative and qualitative evaluation studies are performed on 1235 scans collected from 27 subjects using two different US machines to show comparison results of our model against state-of-the-art GANs for the task of bone surface segmentation using U-net.

5.1 Introduction

Segmentation of bone surfaces from intra-operative US data is an important step for US-guided CAOS procedures. Due to the success of deep learning methods in medical image analysis, recent research has focused on the use of convolutional neural networks (CNNs) for accurate, robust, and real-time segmentation of bone surfaces [61, 43]. However, scarcity of data size, due to a lack of standardized data and patient privacy concerns, is a major challenge in applying deep learning methods in the medical imaging field. This is specifically a challenge due to the fact that US is not a standard imaging modality in CAOS and US-guided CAOS procedures are not common. Another limiting factor is the manual data collection procedure: sub-optimal orientation of the US transducer with respect to the imaged bone anatomy will result in the acquisition of low quality bone scans [33].

Increasing the size of existing datasets through data augmentation in order to improve

models' performance is extensively investigated by various researchers [72]. Earlier work has focused on the introduction of hand crafted image transformations such as random rotations, translations, nonlinear deformations. However, such augmentation methods are limited in their ability to mimic real variations and are highly sensitive to the parameter choice [73]. While transfer learning methods [74], that first train on large datasets then fine-tune on smaller datasets achieve state-of-the-art results on natural image datasets, these methods often do not suit medical image data and offer relatively little benefit to performance [74]. This is especially very problematic for bone US data since its very limited compared to larger medical data such as chest X-ray images. This gap in performance is due to the difference between medical images' features and natural images' features. Furthermore, medical images are often 3D, and there is no streamlined way to transfer 2D feature knowledge into 3D feature knowledge. One approach to overcome this problem is by using unsupervised feature extractors that have only been trained on medical images, however, this requires the target network architecture to be similar to the feature extractors' source architecture, which is uncommon. Image generation methods have recently become a popular solution for the challenge of creating large amounts of training data for deep learning [75]. Generative Adversarial Networks (GANs) have been used in diverse contexts such as unsupervised representation learning [71], image-to-image translation [69] and unsupervised domain adaptation of multi-modal medical imaging data [76]. This groundwork of successful research demonstrates GANs' potential for augmenting small datasets of medical images.

In this work, we propose a computational method, based on a GAN architecture specifically designed to (1) produce synthetic B-mode bone US images and (2) generate their corresponding segmented bone surfaces which can be used as labels. Based on [77] and [78], we show that a duality concept can be adopted for such tasks when implemented by two convolutional blocks, referred to as self-projection and self-attention blocks. We have conducted quantitative and qualitative evaluation studies on 1235 scans collected from 27

subjects using two different US machines. Furthermore, we show comparison results of our model against state-of-the-art GANs presented in [71] and [69] for the task of generating B-mode bone US images. We also evaluate bone surface segmentation accuracy using synthesized B-mode bone US images generated by the networks investigated when tested on Ronneberger's et. al. [16] U-net architecture. Our work is the first report for generating simultaneous B-mode bone US data and corresponding segmentation labels which we believe to be a novel contribution in the field of US-guided CAOS.

5.2 Proposed Method

5.2.1 Network Architecture

Our architecture is based on the common GAN layout utilizing two co-existing neural networks; a generator G that generate synthetic samples and a discriminator D which attempts to discriminate between these generated synthetic samples and real ones [68]. The generator network transforms some pure random noise vectors z (typically a Gaussian) sampled from a prior distribution $p_z(z)$ into new samples such that $\mathbf{x} = G(\mathbf{z})$. The generated image x_g is expected to resemble the real images x_r. On the other hand, the discriminator D has both: (1) real samples with distribution $p_r(x)$ as well as (2) generated samples with distribution $p_g(x)$ and its output $y_s = D(\mathbf{x})$. The gradient information is back-propagated from the discriminator to the generator and hence, the generator optimizes its parameters to generate better images. Gradient-based methods have been proposed to train such a GAN as saddle point optimization problem. However, an imbalance between the training of the generator and the discriminator might occur if the Jensen–Shannon (JS) divergence was used [79] and the discriminator will more likely be too strong, which makes the generator weakly-trained. Moreover, the problem of mode collapse would arise when the distribution $p_g(x)$ learned by the generator was based on limited modes of the real samples distribution $p_r(x)$. This results in weak and limited generations of images. The training of our proposed GAN follows the typical optimization problem such that the discriminator D is

trying to maximize and the generator G is trying to minimize the following objective function $\mathcal{L}(D, G)$:

$$\min_{G} \max_{D} \mathcal{L}(D, G) = \underset{x_r \sim p_{r(x)}}{E} [\log D(x, y)] + \underset{z \sim p_{z(z)}}{E} [\log(1 - D(x, z))] ;$$

In our generator architecture design the encoder maps the input image into a low-dimensional latent space, and the decoder maps the latent representation into the original space. It is trained to generate both US images and their corresponding segmentation images. We adopt the duality concept presented by [77] with our generator G and discriminator D both incorporating dual information into account. Therefore, our proposed GAN architecture generates segmentation masks/label in addition to the synthesized B-mode US images. This is achieved by modifying the GAN architecture to use two-channel images. In vivo real B-mode US data was assigned to the first channel and expert bone segmentation was assigned to the second channel. Based on [59] and [78], we also employ a self-projection and self-attention blocks into the GAN model as shown in Figure 5.1. Our input is processed through convolutional blocks, with each block consisting of several convolutional layers. Our projection blocks, denoted as P, we add a 1×1 convolution to the projected input that is fed-forward through a 1×1 convolution, a 3×3 convolution, and another 1×1 convolution with each convolution operation followed by batch normalization and rectified linear unit (ReLU) activation. We also use a stride of 2 convolutions to upsample the feature maps. On the other hand, our self-attention block, denoted as A, consists of a 1×1 convolution (followed by by batch normalization and Leaky ReLU activation) that is (1) multiplied by a transposed 1×1 convoluted replica resulting in an attention map and (2) multiplied by the attention map to generate self-attention feature maps. The self-attention approach helps modeling wider range image regions. With self-attention features, the generator can associate fine details at every location and associate them with similar portions of the image. In addition, the discriminator can now enforce complicated geometric constraints relative to the overall image [78]. The architecture of the generator can be

summarized as:

- **encoder:** $A32$ $P32$ - $A64$ $P64$ - $A128$ $P128$ - $A256$ $P256$ - $A512$ $P512$
- **decoder:** $A512$ $P512$ - $A256$ $P256$ - $A128$ $P128$ - $A64$ $P64$ - $A32$ $P32$

In our discriminator model a two-input $N \times N$ PatchGAN-like discriminator [69] was used to classify $N \times N$ patches of the input image as real or synthetic. Our discriminator architecture consists of five convolutional blocks, with a final convolution is applied to the last layer to map the 1-dimensional output before applying a Sigmoid function. Batch normalization operations were followed by 0.2-slope leaky ReLU. An Adam solver with a 0.0002 learning rate was used and the structure of the discriminator can be expressed as follows:

- **discriminator:** $A32$ $P32$ - $A64$ $P64$ - $A128$ $P128$ - $A256$ $P256$ - $A512$ $P512$

5.3 Experimental Results

5.3.1 Data Acquisition

To conduct our experiments that particularly target the problem of data limitation in the US-guided CAOS field, we have collected 1235 in vivo B-mode US images categorized into four groups of bone structures: radius, femur, spine and tibia. Data were collected upon obtaining the approval of the institutional review board (IRB). Depth settings and image resolutions varied between 3-8 cm, and 0.12-0.19 mm, respectively. All the collected scans were scaled to a standardized size of 256×256 and manually segmented by an expert ultrasonographer. Two imaging devices were used to collect data:

1. Sonix-Touch US machine (Analogic Corporation, Peabody, MA, USA) with a 2D C5-2/60 curvilinear probe and L14-5 linear probe. Using this device we have collected 1000 scans from 23 subjects. 400 scans from the Sonix touch, using random

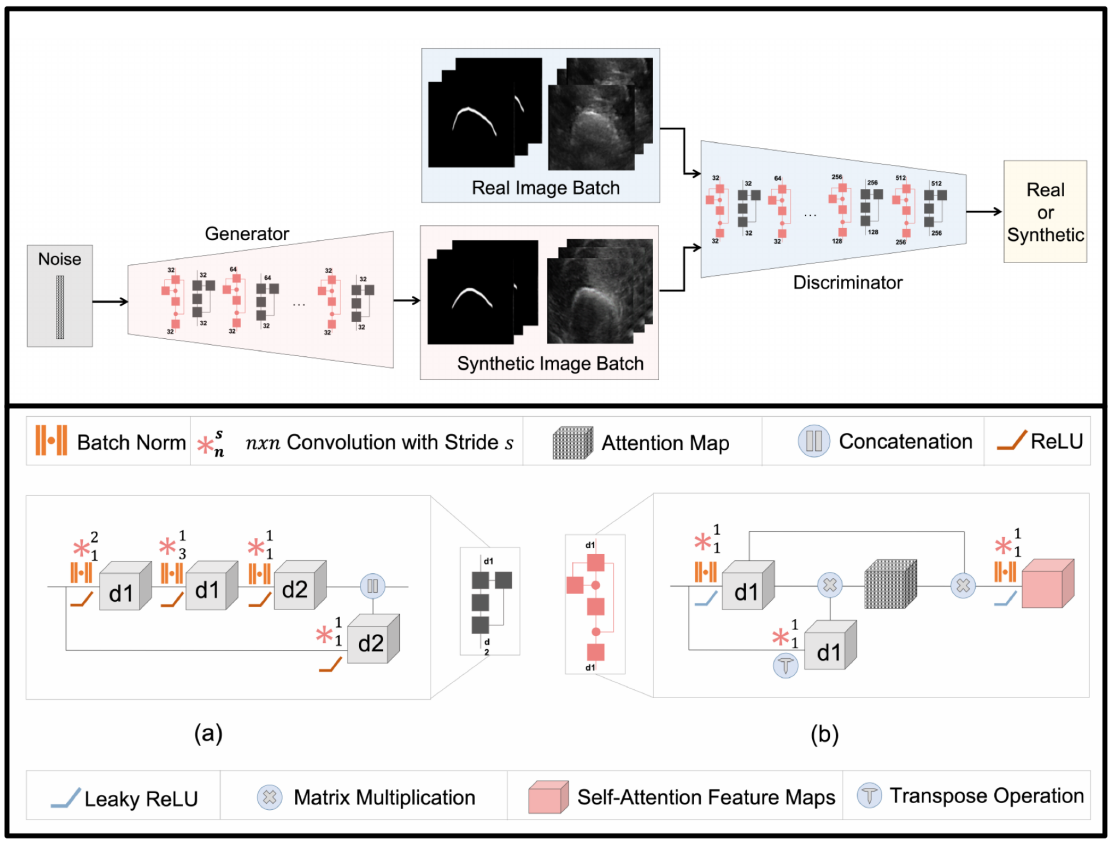

Figure 5.1: Top: An overview of our proposed GAN architecture with its self-projection and attention blocks based generator and patchGAN-like discriminator. Bottom: Our proposed (a) projection blocks in which a 1×1 convolution is concatenated with fed-forward input through a 1×1 convolution, a 3×3 convolution, and another 1×1 convolution with each convolution operation followed by batch normalization and ReLU activation, as presented in [61], and (b) our self-attention blocks in which a 1×1 convolution (with batch normalization and Leaky ReLU activation) is multiplied by a transposed 1×1 convoluted replica resulting in an attention map that is then multiplied by the input to the block to generate self-attention feature maps.

split, were used for training the GANS, 300 scans were used for training the U-net, and 300 scans were used for testing. We repeated this process 3 times and during random split same patient data was not included in the training and testing data.

2. Clarius C3 hand-held wireless ultrasound probe (Clarius Mobile Health Corporation, BC, Canada). Using this device we have collected 235 scans from 4 subjects. All Clarius data was used for testing.

We conducted our experiments using the Keras framework and Tensorflow as backend with an Intel Xeon CPU at 3.00GHz and an Nvidia Titan-X GPU with 8GB of memory. Our GAN converged in about 2 hours during the training process. Testing on average took 35 milliseconds. For our experiments, the proposed network and those presented in [71] and [69] were implemented as per the recommendations by their respective authors. For consistency, we used an Adam solver with learning rate of 0.0002, an exponential decay rate for the first and second moment estimates of $\beta_1 = 0.5$ and $\beta_2 = 0.999$, with a mini-batch SGD for all models considered.

5.3.2 Quantitative Results

Quantitative evaluation of our proposed GAN architecture was performed against three methods [71, 69, 80]. In order to show that the synthesized US images are useful for improving the performance of a supervised segmentation network we use the well known U-net architecture described in [16]. We would like to mention that the U-net architecture used in this work is not the main contribution of this work, since the synthesized images can be used in conjunction with other CNN-based network architectures [61, 43]. If a GAN architecture captures the target distribution correctly it should generate a new set of training images (synthesized images) that should be indistinguishable from the in vivo real B-mode US data. Therefore, a U-net trained on either of these datasets, assuming they have the same size, should produce similar results. To evaluate this we have performed

the following studies: (1) train U-net using limited in vivo real B-mode US data and test using in vivo real B-mode US data, (2) train U-net using limited in vivo real B-mode US data together with synthesized B-mode US data and test using in vivo real B-mode US data, (3) train U-net using synthesized B-mode US data and test U-net using in vivo real B-mode US data, (4) train U-net using real in vivo B-mode US data and test U-net using synthesized B-mode US data. Bone segmentation results are evaluated by calculating Dice, Rand error (Rand), the structural similarity index (SSIM), Hamming Loss, intersection over union (IoU) and average Euclidean distance (AED) [61].

Table 5.1 shows the performance of bone surface segmentation when U-net [16] is trained on various combinations of in vivo real B-mode US data and synthesized B-mode US data. We observe that adding synthesized images to the real in vivo B-mode US images improves the accuracy over the corresponding real-only counterpart. Overall our method outperforms previous state-of-the-art GAN architectures. In particular, it achieves $7\%/7\%$ and $5\%/4\%$ improvement for both data sets (Sonix/Clarius), in IoU value, over the GAN architectures proposed in [71], [69] respectively. A paired t-test, for IoU, Dice and AED results at a $\%5$ significance level, between our proposed network and the networks in [71], [69] achieved p-values less than 0.05 indicating that the improvements of our method are statistically significant. Quantitative results presented in Tables 5.2-5.3 show that our proposed GAN architecture captures the target distribution better compared to the methods in [71, 69] achieving improved results for IoU, Dice, Rand and AED evaluation metrics. Results in Table 5.2 were obtained when U-net [16] was trained using 600 synthetic B-mode US data generated using the proposed and two other architectures [71, 69]. Testing was performed using 300 in vivo real B-mode US data obtained from Sonix Touch and 235 in vivo real B-mode US data obtained from Clarius probe. In Table 5.3 results were obtained when U-net [16] was trained using 535 in vivo real B-mode US data obtained from SonixTouch and Clarius probe. Testing was performed using 600 synthetic B-mode US data generated using the proposed method and two other GAN architectures [71, 69].

5.3.3 Qualitative Results

Qualitative results of our proposed GAN model are shown in Figure 5.2. In each row of Figure 5.2, we demonstrate one example of in vivo real B-mode US image (four examples in total). Columns are labeled alphabetically where we show in (a)-right: real in vivo B-mode US images and in (a)-left: their corresponding bone surface segmentations obtained by an expert. Figure 5.2 columns (b) through (d) demonstrate synthetic B-mode US images (right) and their corresponding synthetic bone surface segmentations as generated by [71], [69] and our proposed model, respectively. Investigating the results we can infer that our proposed method results in fewer artifacts compared to the state-of-the-art [71, 69].

5.4 Discussion and Conclusion

In this paper, a novel GAN model for real-time and accurate B-mode bone US image generation is proposed. Our model has been implemented using two main components: (1) a generator that produces synthesized B-mode US as well as bone surface images and (2) a PatchGAN-like discriminator [69] that was used to classify $N \times N$ patches of the input images as real or synthetic. We have employed two integral components of building the generator and discriminator: a self-projection and self-attention blocks. With self-attention features the generator can associate fine details at every location and associate them with similar portions of the image. The main benefit of our self-attention blocks is that they leverage complementary features in distant portions of the image rather than local regions of fixed shape especially for images with complex structural patterns, e. g. US B-mode images. The relationship between near and far pixels is learned, which allows the model to focus on separated structurally relevant features. Since the task is to replicate the relationship between the US B-mode and segmentation images, our model's ability to span a larger region in the image to create features gives it an advantage over the classic GAN model, which is limited by its filter size. In a classic GAN model, the relationship be-

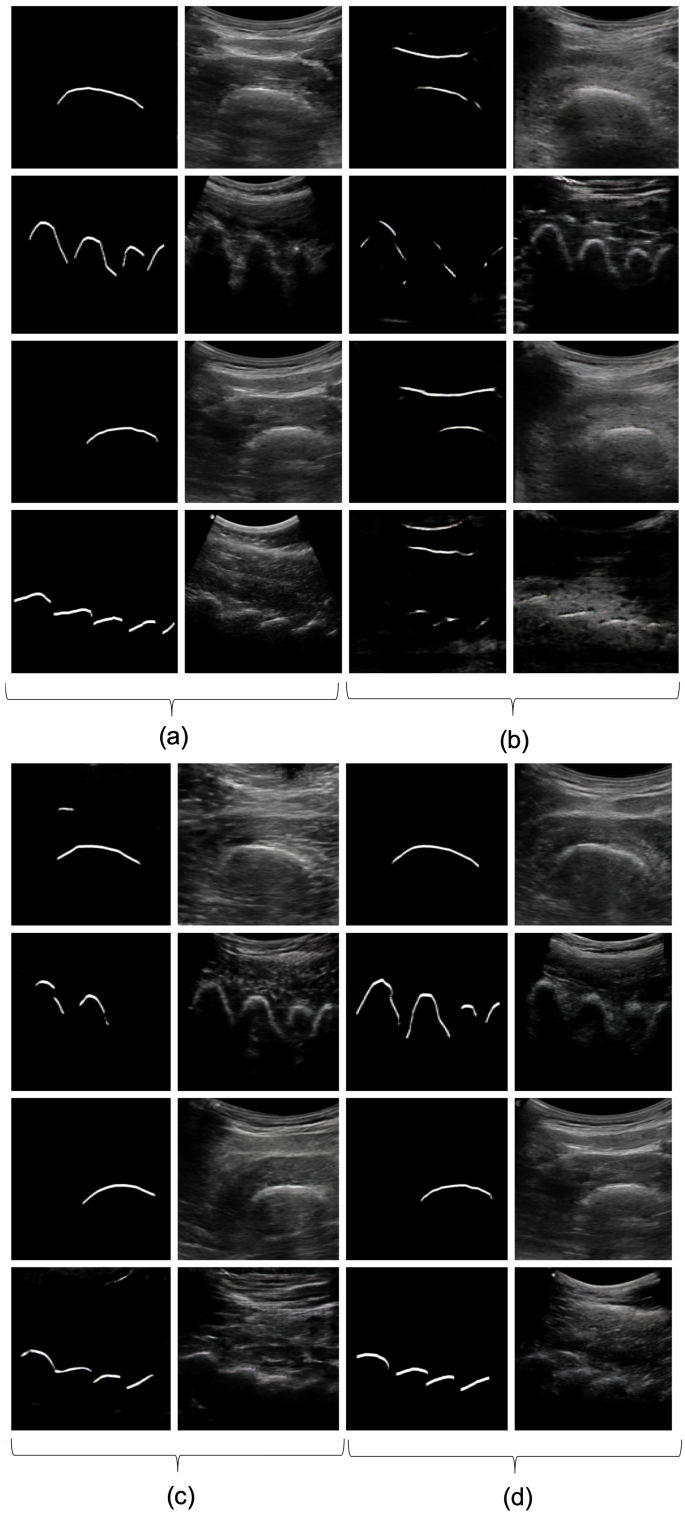

Figure 5.2: Four examples of B-mode US images and their corresponding bone segmentation mask images. (a)-right: real in vivo B-mode US images and in (a)-left: their corresponding bone surface segmentations mask as obtained by an expert. Columns (b) through (d) demonstrate synthetic B-mode US images (right) and their corresponding synthetic bone surface segmentations as generated by [71], [69] and our proposed model.

tween the segment and US features is likely to be diluted across local features, while in a self-attention model the relationship is preserved by these larger feature regions. Additionally, the self-attention discriminator used checks for consistency in features in distant areas, which enforces accurate reproduction of geometric patterns in the B-mode US images and leads to higher-quality augmented data. On the other hand, self-projection blocks allow semantic information to be more efficiently passed forward in the network while progressively increasing feature map sizes, compared to simple convolutions. They allow us to have more comprehensive feature maps. Furthermore, self-projection blocks are also convolutional blocks, and therefore are computationally less expensive to train and infer on. To the best of our knowledge, this was not previously investigated for generating B-mode bone US images. Based on the quantitative results presented, we can conclude that having a self-attention mechanism can significantly improve the results for the image synthesis task at hand. Our future work will involve more extensive clinical validation of the proposed GAN model.

Table 5.1: Quantitative results for bone surface segmentation using U-net [16]. Testing was done using 300 in vivo real B-mode US data obtained from Sonix Touch for Dataset I. For Dataset II testing was performed using all the 235 scans collected from Clarius C3 US probe. Notation note: number of in vivo real B-mode US images/number of synthetic B-mode US images used for training- GAN method used.

Method	IoU%	Dice	Rand	SSIM	Hamming	AED
Dataset I - Sonix-Touch US						
300/0 - N/A	0.7703	0.8642	0.9264	0.1106	0.2280	0.9386
300/300 - Radford et. al. [71]	0.8391	0.9036	0.8522	0.3588	0.1608	0.8146
300/300 - Isola et. al. [69]	0.8516	0.9117	0.8477	0.5401	0.1483	0.5687
300/300 - Ours	**0.8977**	**0.9400**	**0.7899**	**0.7038**	**0.1022**	**0.2985**
300/600 - Radford et. al. [71]	0.8621	0.9183	0.7084	0.5540	0.3826	0.1378
300/600 - Arjovsky et. al. [80]	0.8827	0.9255	0.6876	0.6038	0.1244	0.3220
300/600 - Isola et. al. [69]	0.8943	0.9395	0.6657	0.7021	0.1035	0.2896
300/600 - Ours	**0.9309**	**0.9580**	**0.6125**	**0.7586**	**0.0690**	**0.1596**
Dataset II - Clarius C3 US						
300/0 - N/A	0.7594	0.8564	0.9350	0.1086	0.2405	0.7821
300/300 - Radford et. al. [71]	0.8128	0.8869	0.8678	0.2750	0.1871	0.7536
300/300 - Isola et. al. [69]	0.8322	0.9126	0.8463	0.3483	0.1593	0.8211
300/300 - Ours	**0.8753**	**0.9193**	**0.8381**	**0.5861**	**0.1278**	**0.1970**
300/600 - Radford et. al. [71]	0.8458	0.9128	0.8483	0.4822	0.1486	0.6217
300/600 - Arjovsky et. al. [80]	0.8531	0.9196	0.8104	0.5480	0.1311	0.4853
300/600 - Isola et. al. [69]	0.8646	0.9214	0.7903	0.5728	0.1275	0.3482
300/600 - Ours	**0.9225**	**0.9536**	**0.7636**	**0.7408**	**0.0774**	**0.1583**

Table 5.2: Quantitative results for bone surface segmentation. Results were obtained when U-net [16] was trained using 600 synthetic B-mode US data generated using the proposed method and [71, 69]. Testing was performed using 300 in vivo real B-mode US data (Sonix Touch) and 235 in vivo real B-mode US data (Clarius probe). Notation note: method used-blocks type.

Method	IoU%	Dice	Rand	Hamming	AED
Radford et. al. [71]	0.8471	0.9158	0.8483	0.1783	0.7133
Isola et. al. [69]	0.8625	0.9115	0.8284	0.1183	0.4540
Ours-none	0.6952	0.8068	0.9845	0.1967	0.9347
Ours-self-projection only	0.8356	0.9023	0.8615	0.1883	0.8053
Ours-self-attention only	0.8502	0.9104	0.8816	0.1668	0.5063
Ours-self-projection & self-attention	**0.9054**	**0.9766**	**0.8169**	**0.1208**	**0.1852**

Table 5.3: Quantitative results for bone surface segmentation. Results were obtained when U-net [16] was trained using 535 in vivo real B-mode US data obtained from Sonix Touch and Clarius probe. Testing was performed using 600 synthetic B-mode US data generated using the proposed method and two other GAN architectures [71, 69]. Notation note: number of synthetic B-mode images used for testing - method used.

Method	IoU%	Dice	Rand	Hamming	AED
600-B-mode-Radford et. al. [71]	0.8726	0.9158	0.8464	0.1405	0.4610
600-B-mode-Isola et. al. [69]	0.8933	0.9304	0.7629	0.1108	0.2814
600-B-mode-Ours	**0.9357**	**0.9640**	**0.7195**	**0.0496**	**0.1952**

CHAPTER 6
CONCLUSION AND FUTURE WORK

6.1 Conclusion

In this work, we have developed robust deep learning-based methods that can allow automatic and real-time extraction of bone surfaces in CAOS surgeries. The main contributions and the recommended future work of this thesis can be summarized as follows:

- Our multimodal CNN approach utilized fusion of feature maps and multimodal images to abate sensitivity to variations that are caused by imaging artifacts and low intensity bone boundaries. Our multimodal inputs consisted of B-mode US images and their corresponding local phase filtered counterparts. Fusion operations were investigated for our proposed network using different fusion architectures.

- We improved the multimodal CNN architecture using convolutional blocks (convolutional/projection) instead of convolutional layers. Our projection blocks allow (1) semantic information to be more efficiently passed forward in the network while progressively increasing feature map sizes, and (2) they also allow us to have more comprehensive feature maps.

- We have investigated how to combine information from local phase images and B-mode US data by analyzing different fusion strategies. Our results demonstrated that for the task of bone segmentation fusing B-mode US and local phase features at a later stage outperforms early and mid fusion, specifically for the dataset obtained from Clarius C3 US probe.

- Local phase image features enhanced the bone surface response in the US data, and therefore the B-mode US data and local phase image features were less correlated

in the low level features. The proposed late level fusion network modeled the correlations and interactions between high level features of each modality which outperformed the other fusion networks. A similar investigation was observed with a U-net network late fusion design.

- Our filter layer guided CNN method was quantitatively and qualitatively evaluated on 546 in vivo scans by scanning 14 healthy subjects. We achieved an average F-score above 95% with an average bone surface localization error of 0.2 mm. The reported results are statistically significant compared to state-of-the-art.

- One of the drawbacks of the proposed work is the computational time required for the extraction of local phase image features. This takes on average 1 second (MATLAB implementation) which needs to be improved for real-time CAOS procedures where US is used as an intra-operative imaging modality. During this work the expert manual segmentation was performed by a single expert user. The effect of intra- and inter-user expert bone segmentation on the segmentation results is also crucial.

- Based on a novel GAN architecture, we presented a computational method to segment bone shadow images from in vivo US scans in real-time. We showed how these segmented shadow images can be incorporated, as a proxy, to a multi-feature guided CNN architecture for real-time and accurate bone surface segmentation. Quantitative and qualitative evaluation studies are performed on 1235 scans collected from 27 subjects using two different US machines.

- We provided qualitative and quantitative comparison results against state-of-the-art GANs. We have obtained mean dice coefficient ($\pm$ standard deviation) of 93% ($\pm$ 0.02) for bone shadow segmentation, showing that the method was in close range with manual expert annotation. Statistical significant improvements against state-of-the-art GAN methods (paired t-test p ¡ 0:05) was also obtained. One of the advantages of the proposed work is that bone shadow features are obtained instantaneously

making the computational time required suitable for real-time applications.

- We have also proposes a GAN-based computational method to (1) produce synthetic B-mode US images and (2) their corresponding segmented bone surface masks in real-time. We showed how a duality concept can be implemented for such tasks. Armed by two convolutional blocks, referred to as self-projection and self-attention blocks, our proposed GAN model synthesized realistic B-mode bone US image and segmented bone masks. Quantitative and qualitative evaluation studies were performed on 1235 scans collected from 27 subjects using two different US machines to show comparison results of our model against state-of-the-art GANs for the task of bone surface segmentation using U-net [16].

- With self-attention features the generator can associate fine details at every location and associate them with similar portions of the image. The main benefit of our self-attention blocks is that they leverage complementary features in distant portions of the image rather than local regions of fixed shape especially for images with complex structural patterns, e. g. US B-mode images. The relationship between near and far pixels is learned, which allows the model to focus on separated structurally relevant features. Since the task is to replicate the relationship between the US B-mode and segmentation images, our model's ability to span a larger region in the image to create features gives it an advantage over the classic GAN model, which is limited by its filter size. In a classic GAN model, the relationship between the segment and US features is likely to be diluted across local features, while in a self-attention model the relationship is preserved by these larger feature regions.

6.2 Future Work

While this research work, which detailed robust, accurate, real-time and automatic image segmentation and localization methods for bone structures in US-guided interventional

procedures, achieved its targeted and intended results, some further improvements could be implemented to the proposed methods. Along with recommended clinical validation studies, the following list summarizes the suggested future direction of work:

- For our filter layer guided CNN design future work will involve extensive clinical validation of the proposed method. In particular, this deep learning-based approach could be used as a disease assessment tool with proper clinical validation on US scans collected from patients who are scheduled for orthopedic procedures. As a starting point, one can investigate the accuracy of this model when trained to segment fractured bone regions. In addition, an extension of our network architecture to process volumetric US data. [33].

- More work can be done to reduce the computational time required for the extraction of local phase image features, which takes 1 second on average (MATLAB implementation). For real-time CAOS procedures where US is used as an intra-operative imaging modality, one can explore the possibility of generating local phase images by utilizing the GAN-based approach presented in Chapter 5. The model in Chapter 5 could be trained to generate synthetic US images and their corresponding LP images. Once validated, this solution can resolve the LP images computational time problem.

- The method discussed in Chapter 5 can also be investigated to generate synthesized computed tomography (CT) looking data from US images. Amongst many applications, this can be used as input to a US-CT registration method detailed in [81].

Appendices

RUTGERS

Office of Research and Regulatory Affairs
Arts and Sciences IRB
Rutgers, The State University of New Jersey
335 George Street / Liberty Plaza / Suite 3200
New Brunswick, NJ 08901

orra.rutgers.edu/artsci
732-235-2866

March 28, 2019

Ilker Hacihaliloglu
599 Taylor Rd
Piscataway NJ 08854

P.I. Name: Hacihaliloglu
Protocol #: 15-661M

Dear Ilker Hacihaliloglu:

Initial	Amendment	Continuation	Continuation w/ Amend ✓	Adverse Event

Protocol Title: "Computer Assisted Orthopedic Surgery Using 3D Ultrasound Imaging"

This is to advise you that the above-referenced study has been presented to the Institutional Review Board for the Protection of Human Subjects in Research, and the following action was taken subject to the conditions and explanations provided below:

Approval Date: 3/25/2019 **Expiration Date:** 3/24/2020
Expedited Category(s): 1,4 **Approved # of Subject(s):** 100 **Currently Enrolled:** 29

This approval is based on the assumption that the materials you submitted to the Office of Research and Sponsored Programs (ORSP) contain a complete and accurate description of the ways in which human subjects are involved in your research. The following conditions apply:

- **This Approval-**The research will be conducted according to the most recent version of the protocol that was submitted. **This approval is valid ONLY for the dates listed above;**
- **Reporting-**Reporting- Reporting—ORRA/Arts & Sciences IRB must be immediately informed of any injuries to subjects that occur (within 24 hours) and/or problems (e.g., subject complaints) that arise, in the course of your research within a timely manner (within 5 business days). Visit our website for more information on reportable events, https://orra.rutgers.edu/reportable-events.
- **Modifications-**Any proposed changes MUST be submitted to the IRB as an amendment for review and approval prior to implementation;
- **Consent Form(s)-**Each person who signs a consent document will be given a copy of that document, if you are using such documents in your research. The Principal Investigator must retain all signed documents for at least three years after the conclusion of the research;
- **Continuing Review-**You should receive a courtesy e-mail renewal notice for a Request for Continuing Review before the expiration of this project's approval. However, it is your responsibility to ensure that an application for continuing review has been submitted to the IRB for review and approval prior to the expiration date to extend the approval period;

Additional Notes:
- **Continuation with Amendment Expedited Approval per 45 CFR 46.110(b)(2) for removing study personnel: Prajna Desai and Cosmas Mwikirize**

Failure to comply with these conditions will result in withdrawal of this approval.
Please note that the IRB has the authority to observe, or have a third party observe, the consent process or the research itself. The Federal-wide Assurance (FWA) number for the Rutgers University IRB is FWA00003913; this number may be requested on funding applications or by collaborators.

Respectfully yours,

Acting For--
Beverly Tepper, Ph.D.
Professor, Department of Food Science
IRB Chair, Arts and Sciences Institutional Review Board
Rutgers, The State University of New Jersey

(MW:gj)

RUTGERS
THE STATE UNIVERSITY OF NEW JERSEY

RE: Participation request for Computer Assisted Orthopedic Surgery Using 3D Ultrasound Imaging Study

Rutgers University, School of Engineering, Department of Biomedical Engineering is conducting a study to demonstrate the clinical feasibility of using ultrasound for imaging bone surfaces in orthopaedic surgery applications. The main imaging modality which is going to be used in this study is Ultrasound. Ultrasound is a well-established imaging modality in medicine, which has no known risks and involves no ionizing radiation.

Subjects who are Biomedical Engineering students are invited to participate in this study. The study will involve placing gel over the anatomical area (femur, tibia, knee, pelvis and spine) area and scanning with Ultrasound probe from different angles.

We would appreciate your participation in this study. If you are interested in participating or for more information, please contact Dr. Ilker Hacihaliloglu at +1 848-445-6564 or ilker.hac@rutgers.edu .

Yours sincerely,

Dr. Ilker Hacihaliloglu
Principal Investigator
Dept of Biomedical Engineering

APPROVED
MAR 25 2019
Approved by Rutgers IRB

EXPIRES
MAR 24 2020
Approved by Rutgers IRB

RUTGERS
THE STATE UNIVERSITY OF NEW JERSEY

INFORMED CONSENT FORM

You are invited to participate in a research study that is being conducted by Ilker Hacihaliloglu, who is a professor in the Department of Biomedical Engineering, School of Engineering at Rutgers University. The purpose of this research is to determine the effectiveness and demonstrate the feasibility of using ultrasound, an imaging modality that involves no ionizing radiation, in orthopedic surgery applications for imaging bone. This will be achieve by performing three dimensional (3D) ultrasound imaging of the bone surface (femur, tibia, knee, pelvis and spine) and developing new computational methods that will extract the bone surface from these scans.

Approximately 100 subjects will participate in the study, and each individual's participation will last approximately 20 minutes.

The study procedures include:

- Anatomical areas to be scanned include: Femur, tibia, knee, spine and pelvis bones.
- Scanning time is expected to be 2 minutes per area of interest. Considering the need for repositioning of the ultrasound probe we expect this will take 4 minutes per anatomical area scan. The total scanning time will be limited to 20 minutes.
- Scanning will be performed in the Computer Assisted Surgery and Therapy laboratory (BME 017) located inside the Department of Biomedical Engineering building at Rutgers University.
- Before starting the ultrasound scan a special ultrasound gel (the amount equivalent to the tip of a thumb) will be spread into the skin surface. The gel makes the probe movement much easier and effective. It also helps coupling the probe interface with the skin surface which makes the image quality much better. The gel has no perfumes, no color, is hypoallergenic and is water-soluble.
- The obtained ultrasound volumes will be archived on a secure storage disk for further investigation. The computers that will be used during the data analysis are password protected. Furthermore, the computers that will be used for data analysis are in locked rooms.
- The collected volumes will be stored using a UNIQUE subject number, not derived from personal identifiers-such as spaces/fields for subject name, the first or last three letters of a subject's name, actual initials, reversed initials, birth date, hospital medical record number, provincial personal health number, social insurance number, address or phone number-, will be used. Patients will be identified by the date of scan, gender and age.
- Data sharing: By participating, you understand and agree that the data and information gathered during this study may be used by Department of Biomedical Engineering, Rutgers University and published and/or disclosed by Rutgers University to others outside of Rutgers University. The de-identified data will be shared on the web for research collaboration purposes. However, no personally identifying information will ever be mentioned in any such publication or dissemination of the research data and/or results to other researchers.

RUTGERS
THE STATE UNIVERSITY OF NEW JERSEY

- Subjects who have not provided signed consent and subjects who are: pregnant women, neonates, children, minors, cognitively impaired persons, inmates, elderly, non-english speaking persons will not be included in this study.

The research team and the Institutional Review Board at Rutgers University are the only parties that will be allowed to see the data, except as may be required by law. If a report of this study is published, or the results are presented at a professional conference, only group results will be stated. All study data will be kept for three years. After the study is finished and the results are analyzed and represented in scientific meetings the obtained data will be deleted securely.

The risks of participation include: As this study does not influence any treatment and uses technology which is commonly used in other fields of medicine we foresee no additional risks from participating in this study.

There are no direct benefits in taking part in this study. We hope that the information learned from this study will help in developing 3D ultrasound imaging based image guided surgery system for orthopaedic surgery applications and will be a potential benefit to future studies.

Participation in this study is voluntary. You may choose not to participate, and you may withdraw at any time during the study procedures without any penalty to you. In addition, you may choose not to answer any questions with which you are not comfortable.

If you have any questions about the study or study procedures, you may contact myself at:

Ilker Hacihaliloglu, Ph.D.
Department of Biomedical Engineering
Rutgers University
599 Taylor Road – Room 214
Piscataway, NJ 08854
Phone: +1 848-445-6564
Fax: +1 732-445-3753
Email: ilker.hac@rutgers.edu

If you have any questions about your rights as a research subject, please contact an IRB Administrator at the Rutgers University, Arts and Sciences IRB:

Institutional Review Board
Rutgers University, the State University of New Jersey
Liberty Plaza / Suite 3200
335 George Street, 3rd Floor
New Brunswick, NJ 08901
Phone: 732-235-2866
Email: human-subjects@ored.rutgers.edu

RUTGERS
THE STATE UNIVERSITY OF NEW JERSEY

You will be given a copy of this consent form for your records.

Sign below if you agree to participate in this research study:

Subject (Print) _____

Subject Signature _____ Date _____

Principal Investigator Signature _____ Date _____

For IRB Use Only. This Section Must be Included on the Consent Form and Cannot Be Altered Except For Updates to the Version Date.

IRB Stamp Box
APPROVED
MAR 25 2019
Approved by Rutgers IRB

IRB Stamp Box
EXPIRES
MAR 24 2020
Approved by Rutgers IRB

RUTGERS
THE STATE UNIVERSITY OF NEW JERSEY

Computer Assisted Orthopedic Surgery Using 3D Ultrasound

Principle Investigator: Ilker Hacihaliloglu, Ph.D.

Site(s) where study will be performed:
Department of Biomedical Engineering, School of Engineering, Rutgers University

1. Purpose/Specific Aims

The purpose of this study is to determine the effectiveness and demonstrate the feasibility of using ultrasound, an imaging modality that involves no ionizing radiation, in orthopedic surgery applications for imaging bone. We will achieve this by performing three dimensional (3D) ultrasound imaging of the bone surface (femur, tibia, knee, pelvis and spine) and developing new computational methods that will extract the bone surface from these scans. The results obtained from this study will help during the development of an ultrasound based computer assisted orthopedic surgery system.

Objectives

This research will introduce the concept of using radiation-free real time 3D ultrasound imaging modality for extraction of bone surfaces from volumetric ultrasound images using state of the art image processing methods. Furthermore, it will be invaluable for all future imaging studies with ultrasound in orthopaedic surgeries.

Hypothesis

Our hypothesis is that 3D real-time ultrasound can provide useful information about the bone surfaces for orthopaedic surgery applications.

2. Background and Rationale

The primary medical imaging modalities used in orthopaedic surgeries are X-ray-based radiography, fluoroscopy and computed tomography (CT). Although these modalities typically provide high quality visualization of bone structures, they nevertheless pose several challenges. For example, two dimensional (2D) X-ray data limits the surgeon's ability to visualize the 3D structure of the bone surface which reduces their ability to accurately reduce fractures and safely place implants. Multiple fluoroscopy scans from different views are thus typically required to properly assess bone reduction and to guide the surgical tools and implants. 3D CT images, on the other hand, offer excellent visualization of the imaged anatomic area at high resolutions; however, CT imaging can normally only be acquired either pre- or postoperatively and so is not useful for real-time guidance. Finally, all these imaging operate with ionizing radiation which raises important safety concern to the patient and surgical team. Ultrasound has traditionally been used to image the body's soft tissue, organs, and blood flow in real time. Since there is no

clinically reported risk of using ultrasound, it is still regarded as the only safe method to image a fetus. Consequently, in order to eliminate the substantial exposure of ionizing radiation to both the surgical teams and patients, special attention has been recently given to incorporating ultrasound imaging instead of fluoroscopy, in computer assisted othopaedic surgery systems, which completely eliminates the exposure of ionizing radiation [1-9]. Although unable to penetrate bone, ultrasound strategy is capable of delineating the surface of bone in in two and three dimensions, which may be used as an anatomical landmark [10-12].

3. Design and Methods

The components of the proposed research method are an ultrasound scanner and an ultrasound probe. By interacting with the ultrasound probe the user can acquire 3D ultrasound scans. The acquisition time for one ultrasound volume is approximately 20 seconds, translating into 2 minutes (6 volumes) to assess a single area (femur, tibia, knee, spine or pelvis), and doubling to 4 minutes per area scanned to factor in positioning time of the ultrasound probe.

Due to the physics of ultrasound imaging ultrasound signals cannot penetrate the bone surface. Therefore, only the bone surface which is perpendicular to the probe surface can be scanned and visualized. In order to span all the bone area different ultrasound volumes from different directions, of the same anatomical area, will be obtained. As stated, actual scanning time is expected to be 2 minutes per area of interest. Considering the need for repositioning we expect this will take 4 minutes per scan. As we will need to scan each subject from different anatomical areas (femur, tibia, knee, spine and pelvis) over the course of the study we expect the total scanning time equal to 20 minutes for the mentioned 5 different anatomical areas.

Ultrasound scans will be obtained from subject who are Department of Biomedical Engineering, Rutgers University graduate and undergraduate student. Enrollment in this pilot study will be limited to 100 subjects who have provided informed consent. The subjects will be included in the study after voluntary consent is obtained. The actually scanning time will be scheduled based on the availability of the subject.

US technique: Before starting the ultrasound scan a special ultrasound gel (the amount equivalent to the tip of a thumb) will be spread into the skin surface. The gel makes the probe movement much easier and effective. It also helps coupling the probe interface with the skin surface which makes the image quality much better. The gel has no perfumes, no color, is hypoallergenic and is water-soluble. During the scanning no additional pain will be caused to the patient and the total scanning time will be limited to 20 minutes. The acquired ultrasound volumes will be analyzed with state of the art image processing techniques after transferring them to a computer workstation. The software program that will be used to develop the computational image analysis method is MATLAB (The Mathworks Inc., Natick, MA). MATLAB is a high-level technical computing language and interactive environment for algorithm development, data visualization, data analysis. It is used in most Universities for algorithm development.

Analysis: The ultrasound scans will be transferred to a workstation (PC) located at Computer Assisted Surgery and Therapy Laboratory, a multidisciplinary research laboratory located at the Department of Biomedical Engineering, Rutgers University. We will extract the 3D bone surfaces from the US scans by using state of the art image

processing techniques that are already being developed by the PI Ilker Hacihaliloglu [10-12].

Duration of study
The study will take three years to complete or will terminate after total number of subject (sample size: 100) have been scanned (whichever occurs first).

Study Site
Computer Assisted Surgery and Therapy Laboratory, Department of Biomedical Engineering, Rutgers University (New Brunswick Campus).

Sample Size
The sample size will be limited to 100 subjects in order to provide a statistical analysis.

Subject Selection

Inclusion Criteria
Undergraduate and graduate students of Department of Biomedical Engineering at Rutgers University New Brunswick campus who provided informed consent.

Exclusion Criteria
Subjects who are not able provide informed consent, pregnant women, cognitively impaired persons, non-English speaking persons. Students who are enrolled in PI's class, at the undergraduate and graduate level, will also not be included in the study.

4. Study Variables

Not applicable.

Chart Review Selection
Not applicable.

Risks
As this study does not influence any treatment and uses technology which is commonly used in other fields of medicine we foresee no additional risks from participating in this study. All information gathered will be stored on password protected computers without using any personal identifiers.

Benefits

There are no direct benefits for participating in this study. We hope that the information learned from this study will help in developing 3D ultrasound imaging based image guided surgery system for orthopaedic surgery applications and will be a potential benefit to future studies.

5. Subject Recruitment
The subjects will be recruited using advertisement posted in the Department of Biomedical Engineering building. We will also send a recruitment email to the Department of Biomedical Engineering undergraduate and graduate students email

list notifying them about the study. The subject will be invited to participate and will confirm his/her willingness by reviewing and signing a study participant consent form of which he/she will keep a copy. Subjects who have not provided signed consent and subjects who are: pregnant women, neonates, children, minors, cognitively impaired persons, inmates, elderly, non-english speaking persons will not be included in this study.

Consent Procedures
Written informed consent will be obtained from each subject at entry into the study. Informed consent is obtained by the following process:

1. The subject will be asked to review the study consent form.
2. The PI will meet with the subject to review the form, to confirm the subject's understanding of the study, and to answer any questions that the subject might have.
3. Once the subject demonstrates understanding of the study and agrees to participate in the study, the consent will be signed in the presence of the PI and a witness.

Subject Costs and Compensation
Not applicable.

6. Data Handling
A UNIQUE subject number, not derived from personal identifiers-such as spaces/fields for subject name, the first or last three letters of a subject's name, actual initials, reversed initials, birth date, hospital medical record number, provincial personal health number, social insurance number, address or phone number-, will be used. Patients will be identified by the date of scan, gender and age. The obtained data will be archived on a secure storage disk for further investigation. The computers that will be used during the data analysis are password protected. Furthermore, the computers that will be used for data analysis are in locked rooms. The de-identified data will be shared on the web for research collaboration purposes. However, no personally identifying information will ever be mentioned in any such publication or dissemination of the research data and/or results to other researchers.

7. Statistical Considerations
Data will be analyzed by extracting the bone surfaces from the ultrasound images using the developed state of the art image processing methods. The software program that will be used to develop the computational image analysis method is MATLAB (The Mathworks Inc., Natick, MA). MATLAB is a high-level technical computing language and interactive environment for algorithm development, data visualization, data analysis. It is used in most Universities for algorithm development. Using image intensity information we will develop the image processing method. The generated data will be a three-dimensional (3D) surface mesh model.

8. Data and Safety Monitoring
The PI or study staff will review all data collection forms on an ongoing basis for data completeness and accuracy as well as protocol compliance. The frequency of data review for this study will occur monthly. Only the researchers listed in the group involved with this research work will be allowed to access the data. All data will be

identified only by the code described earlier. The study would be stopped early if we were unable to obtain adequate images with the scanner or the patient was unable to complete the scan. The data will not be used for clinical decision making.

9. Reporting Results

Individual Results

No individual results are relevant to the subjects of this study and thus none will be reported.

Aggregate Results

No aggregate results are relevant to the subjects of this retrospective chart review study and thus none will be reported.

Professional Reporting

Significant results will be submitted for presentation and publication in the appropriate medical forums, specifically publications, conferences, and forums.

10. References

[1] Chen T.K., Abolmaesumi P., Pichora D.R., Ellis R.E., "A system for ultrasound-guided computer assisted orthopaedic surgery", Journal of Computer Aided Surgery, Vol. 10, pp. 281-292, 2005.

[2] Amin D.V., Kanade T., DiGioia A. M., Jaramaz B., "Ultrasound registration of bone surface for surgical navigation", Computer Aided Surgery, pp.1-16, 2001.

[3] Herring J. L., Dawant B. M., Maurer Jr. C.R., "Surface based registration of CT images to physical space for image guided surgery of the spine: a sensitivity study", IEEE Transactions on Medical Imaging, pp.743-752, 1998.

[4] Kryvanos A., Computer assisted surgery for fracture reduction and deformity correction of the pelvis and long bones, PhD thesis, University of Manheim, Germany, 2002.

[5] Barratt D.C., Penney P.G., Chan S.K., Slomczykowski M., Carter T.J., Edwards P.J., Hawkes D.J., "Self calibrating 3D-ultrasound-based bone registration for minimally invasive orthopaedic surgery", IEEE Transactions on Medical Imaging, Vol. 25, No. 3, March 2006.

[6] Penney G.P., Barratt D.C., Chan C.S.K., Slomczykowski M., Carter T.J., Edwards P.J., Hawkes D.J., "Cadaver validation of intensity-based ultrasound to CT registration " , Medical Image Analysis, Vol 10, Issue 3, June 2006.

[7] Muratore D. M., "Towards an image guided spinal surgical system using three dimensional ultrasound: from concept to clinic", PhD thesis, Vanderbilt University, USA, 2002.

[8] Carrat L., Tonetti J., Lavallee S., "Percutaneous computer assisted iliosacral screwing", Medical Image Computing and Computer Assisted Intervention Conference, pp.84-91, 1998.

[9] Carrat L., Tonetti J., Merloz P., "Percutaneous computer assisted iliosacral screwing: Clinical validation", Proc. Medical Image Computing and Computer Assisted Intervention Conference, pp.1229-1237, 2000.

[10] Hacihaliloglu, I. and Abugharbieh, R. and Hodgson, A. J. and Rohling, R. "Automatic Bone Localization and Fracture Detection from Volumetric Ultrasound Images Using 3D Local Phase Features", Ultrasound in Medicine and Biology (UMB) Volume 38, Pages: 128--144, 2012.

[11] Hacihaliloglu, I. and Abugharbieh, R. and Hodgson, A. J. and Rohling, R. N. "Automatic Adaptive Parameterization in Local Phase Feature-Based Bone Segmentation in Ultrasound", Ultrasound in medicine and biology (UMB) Volume 38, Pages: 1689--1703, 2011

[12] Hacihaliloglu, I. and Abugharbieh, R. and Hodgson, A. J. and Rohling, R. N. "Bone surface localization in ultrasound using image phase-based features", Ultrasound in medicine and biology (UMB) Volume 35, Pages: 1475--1487, 2009.

REFERENCES

[1] W. H. Organization, "The burden of musculoskeletal conditions at the start of the new millenium: Report of a who scientific group," *WHO Technical Report Series*, vol. 919, 2003.

[2] "United states bone and joint initiative: The burden of musculoskeletal diseases in the united states (bmus)," *Rosemont, IL. Available at http://www.boneandjointburden.org. Accessed on (3/13/2018)*, vol. Third Edition, 2014.

[3] G. Zheng and L. P. Nolte, "Computer-assisted orthopedic surgery: Current state and future perspective," *Frontiers in surgery*, vol. 2, p. 66, 2015.

[4] H. Sehmbi and S. Perlas, *Basics of Ultrasound Imaging. In: Regional Nerve Blocks in Anesthesia and Pain Therapy*. Springer, 2015.

[5] I. Hacihaliloglu, "3d ultrasound for orthopedic interventions," in *Intelligent Orthopaedics*, Springer, 2018, pp. 113–129.

[6] C. C. Aggarwal, *Neural Networks and Deep Learning*. Springer, 2018.

[7] V. N. Vapnik, *The Nature of Statistical Learning Theory*. Springer, 1995.

[8] Tin Kam Ho, "Random decision forests," in *Proceedings of 3rd International Conference on Document Analysis and Recognition*, vol. 1, 1995, 278–282 vol.1.

[9] K. Suzuki, "Overview of deep learning in medical imaging," *Radiological physics and technology*, vol. 10, no. 3, pp. 257–273, 2007.

[10] D. Ravì, C. Wong, F. Deligianni, M. Berthelot, J. Andreu-Perez, B. Lo, and G.-Z. Yang, "Deep learning for health informatics," *IEEE journal of biomedical and health informatics*, vol. 21, no. 1, pp. 4–21, 2017.

[11] H. Greenspan, B. van Ginneken, and R. M. Summers, "Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1153–1159, 2016.

[12] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[13] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1520–1528.

[14] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[15] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[16] 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*, Springer, 2015, pp. 234–241.

[17] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers, "Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture," in *Asian Conference on Computer Vision*, Springer, 2016, pp. 213–228.

[18] A. Valada, J. Vertens, A. Dhall, and W. Burgard, "Adapnet: Adaptive semantic segmentation in adverse environmental conditions," in *Robotics and Automation (ICRA), 2017 IEEE International Conference on*, IEEE, 2017, pp. 4644–4651.

[19] R. Jacobs, M. Jordan, S. Nowlan, and G. Hinton, "Adaptive mixtures of local experts," *Neural Computation*, vol. 3, no. 1, pp. 79–87, 1991.

[20] K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biological Cybernetics*, vol. 36, pp. 193–202, 1980.

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds., Curran Associates, Inc., 2012, pp. 1097–1105.

[22] K.-L. Du and M. N. S. Swamy, *Neural Networks and Statistical Learning*. Springer, 2019.

[23] W. Di, A. Bhardwaj, and J. Wei, *Deep Learning Essentials*. Packt Publishing, 2018.

[24] B. Xu, N. Wang, T. Chen, and M. Li, "Empirical evaluation of rectified activations in convolutional network," *CoRR*, vol. abs/1505.00853, 2015. arXiv: 1505.00853.

[25] P. Pandey, "Real-time ultrasound bone segmentation and robust us-ct registration for surgical navigation of pelvic fractures," M.S. thesis, The University of British Columbia, 2018.

[26] T. M. Ecker *et al.*, "Percutaneous Screw Fixation of the Iliosacral Joint: A Case-Based Preoperative Planning Approach Reduces Operating Time and Radiation Exposure," *Injury*, vol. 48, no. 8, pp. 1825–1830, 2017.

[27] K. Cleary and T. M. Peters, "Image-Guided Interventions: Technology Review and Clinical Applications," *Annual Review of Biomedical Engineering*, vol. 12, no. 1, pp. 119–142, 2010.

[28] J. A. Noble and D. Boukerroui, "Ultrasound image segmentation: A survey," *IEEE Transactions on Medical Imaging*, vol. 25, no. 8, pp. 987–1010, 2006.

[29] N. Quader *et al.*, "Towards reliable automatic characterization of neonatal hip dysplasia from 3d ultrasound images," *IEEE Transactions on Medical Imaging*, vol. 25, no. 8, pp. 602–609, 2016.

[30] J. G. Thomas, R. A. Peters, and P. Jeanty, "Automatic segmentation of ultrasound images using morphological operators," *IEEE Transactions on Medical Imaging*, vol. 10, no. 2, pp. 180–186, 1991.

[31] I. Hacihaliloglu *et al.*, "Bone segmentation and fracture detection in ultrasound using 3d local phase features," *Medical Image Computing and Computer-Assisted Intervention*, vol. 11, no. 1, pp. 287–295, 2008.

[32] E. M. A. Anas *et al.*, "Bone enhancement in ultrasound based on 3d local spectrum variation for percutaneous scaphoid fracture fixation," *Medical Image Computing and Computer-Assisted Intervention*, pp. 456–473, 2016.

[33] I. Hacihaliloglu, P. Guy, A. J. Hodgson, and R. Abugharbieh, "Volume-specific parameter optimization of 3d local phase features for improved extraction of bone surfaces in ultrasound," *The International Journal of Medical Robotics and Computer Assisted Surgery*, vol. 10, no. 4, pp. 461–473, 2014.

[34] ——, "Automatic extraction of bone surfaces from 3d ultrasound images in orthopaedic trauma cases," *International journal of computer assisted radiology and surgery*, vol. 10, no. 8, pp. 1279–1287, 2015.

[35] A. K. Jain and R. H. Taylor, "Understanding bone responses in b-mode ultrasound images and automatic bone surface extraction using a bayesian probabilistic framework," *Proceedings of the SPIE*, vol. 5373, pp. 131–142,

[36] P. Foroughi, E. Boctor, M. J. Swartz, R. H. Taylor, and G. Fichtinger, "P6d-2 ultrasound bone segmentation using dynamic programming," in *Ultrasonics Symposium, 2007. IEEE*, IEEE, 2007, pp. 2523–2526.

[37] A. Karamalis, W. Wein, T. Klein, and N. Navab, "Ultrasound confidence maps using random walks," *Medical Image Analysis*, vol. 16, no. 6, pp. 1101–1112, 2012.

[38] N. Quader, A. Hodgson, and R. Abugharbieh, "Confidence weighted local phase features for robust bone surface segmentation in ultrasound," in *Workshop on Clinical Image-Based Procedures*, Springer, 2014, pp. 76–83.

[39] lEO Grady *et al.*, "Random walks for interactive organ segmentation in two and three dimensions: Implementation and validation," *Medical Image Computing and Computer-Assisted Intervention*, vol. 8, pp. 773–780, 2005.

[40] I. Hacihaliloglu, "Localization of bone surfaces from ultrasound data using local phase information and signal transmission maps," in *International Workshop and Challenge on Computational Methods and Clinical Applications in Musculoskeletal Imaging*, Springer, 2017, pp. 1–11.

[41] M. Salehi, R. Prevost, J.-L. Moctezuma, N. Navab, and W. Wein, "Precise ultrasound bone registration with learning-based segmentation and speed of sound calibration," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2017, pp. 682–690.

[42] N. Baka, S. Leenstra, and T. van Walsum, "Ultrasound aided vertebral level localization for lumbar surgery," *IEEE transactions on medical imaging*, vol. 36, no. 10, pp. 2138–2147, 2017.

[43] M Villa, G Dardenne, M Nasan, H Letissier, C Hamitouche, and E Stindel, "Fcn-based approach for the automatic segmentation of bone surfaces in ultrasound images," *International journal of computer assisted radiology and surgery*, vol. 13, no. 11, pp. 1707–1716, 2018.

[44] F. Ozdemir, E. Ozkan, and O. Goksel, "Graphical modeling of ultrasound propagation in tissue for automatic bone segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2016, pp. 256–264.

[45] I. Hacihaliloglu, "Enhancement of bone shadow region using local phase-based ultrasound transmission maps," *International Journal of Computer Assisted Radiology and Surgery*, vol. 12, no. 6, pp. 951–960, 2017.

[46] I. Hacihaliloglu, A. Rasoulian, R. N. Rohling, and P. Abolmaesumi, "Local phase tensor features for 3-d ultrasound to statistical shape+ pose spine model registration," *IEEE transactions on medical imaging*, vol. 33, no. 11, pp. 2167–2179, 2014.

[47] Z. Guo, X. Li, H. Huang, N. Guo, and Q. Li, "Medical image segmentation based on multi-modal convolutional neural network: Study on image fusion schemes," *arXiv preprint arXiv:1711.00049*, 2017.

[48] J. Liu, S. Zhang, S. Wang, and D. N. Metaxas, "Multispectral deep neural networks for pedestrian detection," *arXiv preprint arXiv:1611.02644*, 2016.

[49] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2014, pp. 1725–1732.

[50] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2016, pp. 1933–1914.

[51] I. Hacihaliloglu, "Localization of bone surfaces from ultrasound data using local phase information and signal transmission maps," in *Computational Methods and Clinical Applications in Musculoskeletal Imaging*, Springer International Publishing, 2018, pp. 1–11.

[52] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical association*, vol. 66, no. 336, pp. 846–850, 1971.

[53] V. Jain, B. Bollmann, M. Richardson, D. R. Berger, M. N. Helmstaedter, K. L. Briggman, W. Denk, J. B. Bowden, J. M. Mendenhall, W. C. Abraham, K. M. Harris, N. Kasthuri, K. J. Hayworth, R. Schalek, J. C. Tapia, J. W. Lichtman, and S. H. Seung, "Boundary learning by optimization with topological constraints," in *2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2010, pp. 2488–2495.

[54] C. Cernazanu-Glavan and S. Holban, "Segmentation of bone structure in x-ray images using convolutional neural network," *Adv. Electr. Comput. Eng*, vol. 13, no. 1, pp. 87–94, 2013.

[55] I. Hacihaliloglu, "Ultrasound imaging and segmentation of bone surfaces: A review," *Technology*, vol. 05, no. 02, pp. 74–80, 2017.

[56] R. Jia, S. J. Mellon, S Hansjee, A. Monk, D. W. Murray, and J. A. Noble, "Automatic bone segmentation in ultrasound images using local phase features and dy-

namic programming," in *Biomedical Imaging (ISBI), 2016 IEEE 13th International Symposium on*, IEEE, 2016, pp. 1005–1008.

[57] N. Baka, S. Leenstra, and T. van Walsum, "Random forest-based bone segmentation in ultrasound," *Ultrasound in Medicine and Biology*, vol. 43, no. 10, pp. 2426–2437, 2017.

[58] P. Wang, V. M. Patel, and I. Hacihaliloglu, "Simultaneous segmentation and classification of bone surfaces from ultrasound using a multi-feature guided cnn," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, vol. 11073, Springer, 2018, pp. 134–142.

[59] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, "Deeper depth prediction with fully convolutional residual networks," in *2016 Fourth International Conference on 3D Vision (3DV)*, IEEE, 2016, pp. 239–248.

[60] S. Schumann, "State of the art of ultrasound-based registration in computer assisted orthopedic interventions," in *Computational Radiology for Orthopaedic Interventions*, Springer, 2016, pp. 271–297.

[61] A. Z. Alsinan, V. M. Patel, and I. Hacihaliloglu, "Automatic segmentation of bone surfaces from ultrasound using a filter layer guided cnn," *International Journal of Computer Assisted Radiology and Surgery*, vol. 14, no. 5, pp. 775–783, 2019.

[62] F. Berton, F. Cheriet, M.-C. Miron, and C. Laporte, "Segmentation of the spinous process and its acoustic shadow in vertebral ultrasound images," *Computers in biology and medicine*, vol. 72, pp. 201–211, 2016.

[63] H. El-Hariri, K. Mulpuri, A. Hodgson, and R. Garbi, "Comparative evaluation of hand-engineered and deep-learned features for neonatal hip bone segmentation in ultrasound," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2019, pp. 12–20.

[64] T. Klein and W. M. Wells, "Rf ultrasound distribution-based confidence maps," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, vol. 9350, 2015, pp. 595–602.

[65] P. Hellier, P. Coupé, X. Morandi, and D. L. Collins, "An automatic geometrical and statistical method to detect acoustic shadows in intraoperative ultrasound brain images," *Medical Image Analysis*, vol. 14, no. 2, pp. 195–204, 2010.

[66] R. Hu, R. Singla, F. Deeba, and R. N. Rohling, "Acoustic shadow detection: Study and statistics of b-mode and radiofrequency data," *Ultrasound in medicine & biology*, vol. 45, no. 8, pp. 2248–2257, 2019.

[67] Q. Meng, J. Housden, J. Matthew, D. Rueckert, J. A. Schnabel, B. Kainz, M. Sinclair, V. Zimmer, B. Hou, M. Rajchl, N. Toussaint, O. Oktay, J. Schlemper, and A. Gomez, "Weakly supervised estimation of shadow confidence maps in fetal ultrasound imaging," *IEEE Transactions on medical imaging*, vol. 38, no. 12, pp. 2755–2767, 2019.

[68] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, pp. 2672–2680, 2014.

[69] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2017, pp. 5967–5976.

[70] B. Xu, N. Wang, T. Chen, and M. Li, "Empirical evaluation of rectified activations in convolutional network," *arXiv preprint arXiv:1505.00853*, 2015.

[71] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.

[72] C. Payer, D. Stern, H. Bischof, and M. Urschler, "Regressing heatmaps for multiple landmark localization using cnns," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, ser. Lecture Notes in Computer Science. Switzerland: Springer International Publishing AG, Oct. 2016, vol. 9901, pp. 230–238.

[73] A. Zhao, G. Balakrishnan, F. Durand, J. V. Guttag, and A. V. Dalca, "Data augmentation using learned transformations for one-shot medical image segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019, pp. 8543–8553.

[74] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio, "Transfusion: Understanding transfer learning for medical imaging," in *Advances in Neural Information Processing Systems*, 2019, pp. 3342–3352.

[75] H.-C. Shin, N. A. Tenenholtz, J. K. Rogers, C. G. Schwarz, M. L. Senjem, J. L. Gunter, K. P. Andriole, and M. Michalski, "Medical image synthesis for data augmentation and anonymization using generative adversarial networks," in *International workshop on simulation and synthesis in medical imaging*, Springer, 2018, pp. 1–11.

[76] K. Kamnitsas, C. Baumgartner, C. Ledig, V. F. J. Newcombe, J. P. Simpson, A. D. Kane, D. K. Menon, A. Nori, A. Criminisi, D. Rueckert, and B. Glocker, *Unsuper-*

vised domain adaptation in brain lesion segmentation with adversarial networks, 2016. arXiv: 1612.08894 [cs.CV].

[77] T. Neff, C. Payer, D. Štern, and M. Urschler, "Generative adversarial networks to synthetically augment data for deep learning based image segmentation," May 2018.

[78] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, *Self-attention generative adversarial networks*, 2018. arXiv: 1805.08318 [stat.ML].

[79] A. K. Yadav, S. Shah, Z. Xu, D. W. Jacobs, and T. Goldstein, "Stabilizing adversarial nets with prediction methods," in *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, OpenReview.net, 2018.

[80] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proceedings of the 34th International Conference on Machine Learning*, D. Precup and Y. W. Teh, Eds., ser. Proceedings of Machine Learning Research, vol. 70, International Convention Centre, Sydney, Australia: PMLR, 2017, pp. 214–223.

[81] X. Zeng, M. Vives, and I. Hacihaliloglu, "Hierarchical 3-d registration of computed tomography to ultrasound using reinforcement learning," *EPiC Series in Health Sciences*, vol. 4, pp. 306–311, 2020.