IMPROVED AUTOMATIC BONE SEGMENTATION USING LARGE-SCALE SIMULATED ULTRASOUND DATA TO SEGMENT REAL ULTRASOUND BONE SURFACE DATA

By

HRIDAYI PATEL

A thesis submitted to the

School of Graduate Studies

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Master of Science

Graduate Program in Biomedical Engineering

Written under the direction of

Ilker Hacihaliloglu

And approved by

_____

_____

_____

New Brunswick, New Jersey

May 2020

ABSTRACT OF THE DISSERTATION

Improved Automatic Bone Segmentation Using Large-Scale Simulated Ultrasound Data to Segment Real Ultrasound Bone Surface Data

by HRIDAYI PATEL

Thesis Director:

Ilker Hacihaliloglu

Automatic segmentation of bone surfaces from ultrasound images is of great interest in the ultrasound-guided computer assisted orthopedic surgery field. These automatic segmentations help the system locate where the bone surface is in the image which can allow for proper surgical manipulation. Methods that involve using image processing tools have previously been used to perform the segmentations however, they have faced problems due to the noise and various imaging artifacts associated with ultrasound data. Most recently, methods based on deep learning have achieved promising results. However, a drawback is that these methods require large number of training dataset. Therefore, new methods which can overcome these drawbacks need to be investigated in order to accurately segment bone surfaces from real ultrasound data.

This thesis introduces the concept of training the deep learning methods with large-scale simulated bone ultrasound data and investigating how using large-scale simulated data along with limited real ultrasound data affects the segmentation performance of the deep learning network. A transfer learning approach and using a training dataset consisting of both real and simulated ultrasound bone surface data was applied for the investigation.

We show that by using simulated bone ultrasound data, the success of traditional deep learning methods increases compared to using small-scale real ultrasound data only.

Data used in the study consisted of real ultrasound data collected from different subjects and utilizing 3D Slicer and PLUS for generating simulate ultrasound data. Various networks were trained in order to determine how well the network of a certain dataset is able to perform automatic segmentations on the same type of data. Additionally, networks trained with both large-scale simulated US data and limited real ultrasound data were trained and tested on real ultrasound data to determine if using large-scale simulated data improves network performance. The automatic segmentations of the neural networks were compared against manual segmentations of the same data by calculating the Sorensen- Dice Coefficient and Average Euclidean Distance. Results of the thesis show that using large-scale simulated ultrasound data can be used to train a neural network to segment real ultrasound data if both types of datasets are used together to develop the network.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

Throughout this project, I have received a lot of support from a lot of different people. I would first like to thank Dr. Ilker Hacihaliloglu who has helped me and guided me throughout this whole process. He has taught me a lot about research over the course of the years and I am extremely grateful to have had the chance to work with him and learn from him on this topic. I would additionally like to thank my committee members, Dr. Nada Boustany and Dr. Mark Pierce, for their feedback regarding this project.

Additionally, I would like to thank my parents and sister for the support they have shown me throughout my time at Rutgers and while working on the thesis work. They have provided a lot of encouragement and love during this whole project and have taught me how to try my hardest for any task I set my mind to. I would also like to thank all my friends for their continued support. They have made sure that I was motivated and inspired throughout the process and have always been there.

I would like to dedicate this thesis to my parents who have sacrificed so much to ensure that I have a bright future. Thank you for always believing in me.

CHAPTER 1

INTRODUCTION

1.1 Thesis Motivation

Between 2012 to 2014, the direct and indirect annual cost for musculoskeletal disease was estimated to be $322 billion [1]. One of the most common types of musculoskeletal injuries include falls which account for 51.7% of hospitalizations and 35.7% of emergency department visits [1]. The most frequent injury that is a result of falls is a fracture which makes up for 80% of hospitalizations and 33% of emergency department visits [1]. Fractures also account for 63% of traumatic injuries seen in hospital cases [1]. One of the ways to treat fractures and other musculoskeletal medical conditions includes orthopedic surgery.

As more and more new research and technologies are being explored in the medical device industry, the field of surgery is experiencing many breakthroughs in surgical equipment and computer guided surgery. These breakthroughs help to develop more efficient and safer surgical practices for the patients. Orthopedic surgery in particular is well suited for computer assistance as bones and periarticular tissues can be imaged easily using pre-operative X-rays, computed tomography (CT), magnetic resonance imaging (MRI), and fluoroscopy [2]. The dominant intra-operative imaging modality in computer-assisted orthopedic surgery (CAOS) is 2D/3D fluoroscopy. 2D fluoroscopy is limited to projection imaging which causes difficulties during fixation of complex fractures. On the other hand, 3D fluoroscopy has improved the surgical success rates by providing 3D

guidance [3]. However, 3D fluoroscopy units are expensive and not as widely available as 2D units. Finally, one of the major healthy concerns of intra-operative fluoroscopy is the exposure, of the surgical team and patient, to harmful ionizing radiation.

New research techniques have looked into using the ultrasound (US) in order to perform intra-operative imaging in CAOS for reducing the exposure to ionizing radiation [4]. US images however are noisy, have artifacts, have a limited field of view and bone surfaces appears several millimeters in thickness (***Fig 1.1***) [4]. Due to these limitations, interpretation of US bone images is difficult. In order to overcome these challenges researchers have looked into developing automatic bone segmentation and registration methods. Accurate segmentation is important for improved guidance in US-guided CAOS systems. The segmented bone surfaces are also used for real-time intra-operative registration [5]. Therefore, accurate segmentation is also important for robust intra-operative registration.

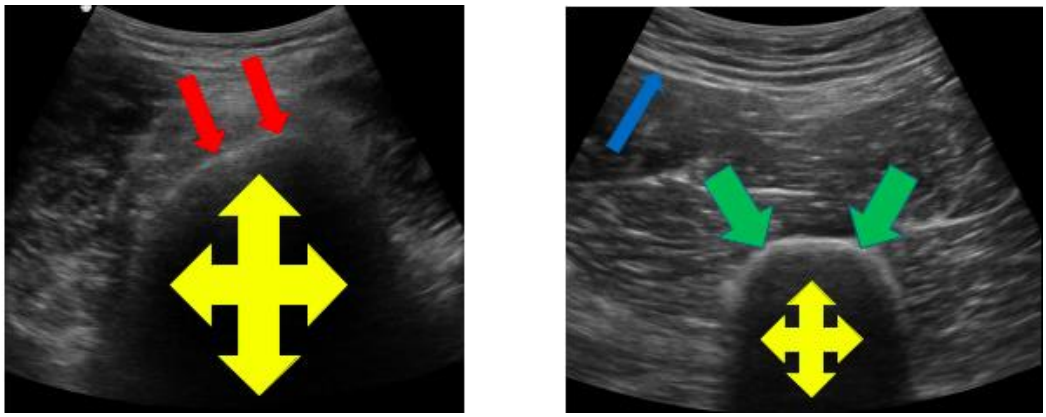


Fig. 1.1 Two B-mode US images of bone surfaces are shown above. The image on the left shows a high-quality US image as the bone surface (red arrows) is shown in high intensity and is followed by a shadow region (yellow arrows). In comparison, the image on the right shows the different artifacts and noise that can occur in a US image. The soft tissue interface (blue arrow) can be misinterpreted as the bone surface (green arrows) by an algorithm. Additionally, noise in the shadow region can occur due to suboptimal probe orientation. This noise may also interfere with the machine learning algorithms segmentation of the bone surface.

Early work for segmenting bone surfaces from US data was based on the use of image intensity or gradient information [6]. However, intensity-based methods are not robust due to the typical imaging artifacts associated with image acquisition or the noise in the US scan which can result in low accuracy segmentation. In order to overcome this, methods based on local phase image information have been developed [7]. Although local phase information based methods can provide accurate and robust segmentation results, their success depends on the optimization of specific filter parameters used to extract phase information. Furthermore, methods based on local phase image information are also time

consuming and not suitable for real-time imaging. Recently, methods based on deep learning have been investigated by various groups for automatic, accurate and real-time segmentation of bone surfaces from US data [8, 9]. However, a limitation of using deep learning methods includes limited data availability for training the neural networks. Two methods, traditionally used by researchers, in order to overcome the scarcity of medical image data are data augmentation and transfer learning. During transfer learning an existing deep learning architecture, designed for natural image datasets, is fine-tuned using the sparse new medical image data [10]. Data augmentation is obtained by introducing random image transformations, rotations or nonlinear deformations, resulting in the generation of new image datasets. Transfer learning and data augmentation have achieved improved results for various tasks such as classification and segmentation [11, 12]. Data augmentation and transfer learning have also been used for various US image analysis methods [13].

Since the US is not the standard imaging modality used in orthopedics, there is a need to efficiently perform segmentations with a smaller dataset. Therefore, this thesis aims to answer the questions:

1) Can simulated US data be used to train a neural network for accurate and robust segmentation of real i*n vivo* bone US data?
2) Can simulated and real *in vivo* bone US data be used together to train a neural network for accurate and robust segmentation of real *in vivo* bone US data?

1.2 Computer Assisted Orthopedic Surgery

Computer Assisted Orthopedic Surgery is the implementation of computer-based technology in order to perform successful orthopedic surgery. The area was introduced in the mid-1990s when the first successful spine and hip replacement surgery was performed using CAOS systems [3]. In orthopedic surgery, precision is of utmost importance because fracture fixation, implant fixation, etc. need to be carried out efficiently in order to minimize post-operation risks [3]. CAOS systems "allows surgeons to get real-time feedback" about the surgical incisions so that they are able to navigate clearly during the surgical procedure [3]. CAOS system components include the preoperative and intraoperative plan, and registration. The preoperative plan is developed when a CT scan of a patient's fracture is taken in order for the surgeon to have a reference image to help determine the proper surgical plan [14]. In order for the CAOS system to help perform the surgery, the patient's anatomy needs to be known during the operation as well. Intraoperative imaging consists of using X-rays, fluoroscopy, or more recently, US imaging [3, 14]. The intraoperative model then needs to be matched with the preoperative model via registration in order to bridge the gap between the preoperative and intraoperative plan [3]. After this registration process is complete, a three dimensional (3D) and real-time image of the patient's bone surface can be obtained which can then be used to perform a more precise and accurate surgery [14].

1.3 Challenges in Computer Assisted Orthopedic Surgery (CAOS)

Due to the importance of imaging in CAOS systems, a proper and safe method of imaging the bone structures needs to be determined. The CT scan is a great imaging tool for use in orthopedic surgery because it provides a 3D image with good contrast between the bone and tissue interface [3]. However, CT imaging is not the ideal method to be used during the intraoperative planning as it requires large changes in the hospital's layout which can come with great financial costs [3]. Therefore, intraoperative imaging uses fluoroscopy to view the bone structures in real-time. However, fluoroscopic images only provide two-dimensional (2D) information which is why several images need to be taken from different planes in order to match them to the 3D CT images [15]. These images require the surgeon to use trial-and-error to determine the placement of the implant in different planes [16]. Although 3D fluoroscopy units are available, they are much more expensive than traditional 2D fluoroscopy which might not make them feasible to obtain [16]. In order to perform the surgery, the surgeon needs to fix the surgical tools and implant relative to the bone. This process relies on the surgeon's expertise and knowledge of anatomical areas from previous surgeries and trainings. 2D fluoroscopy images are relied upon for the surgeon to obtain an accurate estimate of the trajectory to reach the target structure [16].

In addition to the imaging and navigation challenges, a large area of concern is the safety of the patients and surgeons involved in the surgical procedures. In a study performed by Gausden et al. tracked the radiation exposure of surgeons, it was found that orthopedic surgeons in the study received an average of 0.2 to 79 mrem/month (mrem is a unit of radiation exposure) [17]. The senior surgeons who performed multiple trauma surgeries were exposed to more radiation than their peers [17]. Although these results are

under the dose limit that is 5,000 mrem/year, considering the amount of surgeries a surgeon performs it is better to keep radiation exposure lower [17]. The importance of imaging and guidance during surgery can also been seen in the number of radiology images that are requested by the emergency department. Blane et al. showed that 72,886 imaging studies were requested in 2004 from the radiology department [18]. As seen by this number, imaging plays a large role in the diagnosis and treatment of patients coming into the emergency department who possibly require surgery. Therefore, it is important to develop an accurate method of performing surgery that reduces risks to the patients and surgeons, is more efficient, and less invasive.

1.4 US-Based CAOS Systems

The use of the US as an imaging modality has primarily been restricted to imaging soft tissues and internal organs. From 2000-2011, the amount of CTs and MRIs performed have doubled [19]. In comparison, the number of US imaging done has increased tenfold [19]. This can be attributed to the fact that the US is a robust imaging modality which can offer real time feedback and lack of radiation. Due to these advantages, the US can be used as a good replacement for fluoroscopy for intraoperative imaging. The CT and MRI scans can be used as preoperative imaging with the US used as an intraoperative imaging method. Using US based CAOS systems comes with two potential areas of improvement. One area is registering the intraoperative US image to the preoperative CT/MRI image and the second area is the bone segmentation that needs to be done on the US image in order to register it to the preoperative image. The goal of this work is to improve the bone

segmentation accuracy by investigating the effect of using simulated US data on an automatic segmentation process.

1.5 Literature Review of Previous Methods for Bone Surface Segmentations from US

Prior work on segmenting bone surfaces from US data has used both image processing methods and deep learning methods. Kowal et al. used different types of filtering and thresholding including depth-weighted thresholding in combination with contour filters to attempt to segment the bone [6]. They found a difference of an average of 0.42 mm of distance error with a 0.19 mm standard deviation between their automatically segmented contours versus reference contour points [6]. Their segmentation also did not require any manual intervention but as it depended on image morphology, the contrast conditions of the US scans can affect the accuracy of the segmentation [6]. Daanen et al. also evaluated an automated image processing technique in order to segment bone surfaces. They developed “fuzzy intensity images” based on the property that the pixel intensity is highest at the bone surface and used that to develop gradient images [20]. These gradient images were then further processed using wavelet transformations and thresholding to create a segmentation [20]. They found that the mean error for the automated segmentation was less than 10 pixels when compared to the manual segmentations [20]. However, their method also relies on image processing methods which means that the amount of noise in the US image can affect the segmentation which also needs to be taken into account [20]. A 2016 method of bone structure segmentation of US data developed by Jia et al. used “acoustic characteristics of the intensity profile during the US scan to eliminate the soft tissue interference” that occurs in an US image [21].

Additionally, they combined local phase features to determine the areas of the image which have high likelihood of being bone structures. These results were compared to manual segmentations using the average Euclidean distance and they found a very low (0.2 mm) error between the automatic and manual segmentations [21]. However, all of these image processing methods need to incorporate the noise in the US data while retaining the full bone surface profile.

Due to this limitation, neural networks have shown great promise in automatic segmentations [8, 9, 22, 23]. They do not require as much image processing on the US data and rely on a more dataset-based approach for training the neural networks. Villa et al. took into account inter- and intra-user variability while developing a new algorithm based on fully convolutional neural networks (FCN) and compared it to the confidence in phase symmetry (CPS) method [22]. They did this by removing images from the training dataset which had a higher inter-user variability than a confidence threshold value. The FCN based algorithm outperformed the CPS method on all their validation methods except the recall calculation. The average RMSE value for the FCN based method was 1.3mm compared to a 5mm RMSE for CPS methods [22]. Convolutional neural networks (CNN) that incorporate fusion of feature maps and multi-modal images have also been explored by Alsinan et al [23]. Due to the use of real US data for the CNN, variations in the data due to the image artifacts can interfere with a proper bone surface segmentation. Therefore, Alsinan et al. enhanced the bone surface using local phase image feature extraction and used the local phase filtered images and B-mode US images to train a CNN with different fusion layers [23]. The study found that the average Euclidean distance for their late fusion design was 0.1482 mm while for the U-net using normal B-mode US images was 2.296

mm [23]. Wang et al. proposed using a pre-enhancing network along with a modified U-net to segment bone surfaces from real US data. The pre-enhancement network, which uses the B-mode US image and three filtered image features, serves to make the bone surface more dominant in the real US data so that modified U-net can perform the automatic segmentation more effectively [8]. The modified U-net and pre-enhanced network had an average Euclidean distance of 0.246 compared to an average Euclidean distance of 0.435 of using only the U-net [8]. Salehi et al. proposed another bone surface detection method that uses a CNN in order to make registration to pre-operative data more accurate [9]. The CNN used was based off the U-net and was used to generate fuzzy probability maps of the bone surface locations in the US image [9]. Speed of sound calibration was then used to make sure that bone surfaces were not sensitive to the type of tissue present in the US images [9]. The results of the proposed neural network were compared to another feature-based network and the random forest segmentation method [9]. The average dice coefficient was 0.87 for the new neural network compared to 0.44 and 0.79 for the feature-based and random forest methods, respectively [9]. El-Hariri et al. evaluated different image processing methods and the U-net for segmenting hip bone from 439 images of US data [24]. They investigated the shadow peak and confidence-weighted structured phase symmetry methods along with a normal U-net network and a multi-channeled U-net network [24]. The results showed that both types of U-net outperformed the image processing methods and had average dice coefficients of 0.86 and 0.92 both datasets tested [24]. Another study investigated the use of the U-net for automatic segmentation of the spinous process from US data [25]. When results were compared to the random forest

algorithm for segmentation, the U-net outperformed the random forest on the test dataset with an F-score of 0.90 vs 0.83 for random forest [25].

Previous Studies	Methods	Results
Kowal et al. [6]	Combination of filters and depth-weighted thresholding	Average of 0.42 mm distance error between automatically segmented contours vs. reference contour points
Daanen et al. [20]	Development of "fuzzy intensity images" and using gradients and wavelet transformations	Mean error of automatic segmentations was less than 10 pixels when compared to manual segmentations
Jia et al. [21]	Used acoustic characteristics of intensity profile during US scans and local phase features	Automatic segmentations when compared to manual segmentations had 0.2 mm average Euclidean distance
Villa et al. [22]	Accounted inter- and intra-user variability when using FCN for segmentations	Algorithm outperformed CPS method and average RMSE values for FCN based method was 1.3 mm
Alsinan et al. [23]	Used a CNN that incorporated fusion of feature maps and multi-modal images	Average Euclidean distance for late fusion design was 0.1482 mm
Wang et al. [8]	Developed a pre-enhancement network and then used a modified U-net to segment bone surfaces	Modified U-net and pre-enhanced network had an average Euclidean distance of 0.246
Salehi et al. [9]	Incorporated speed of sound calibration after using U-net to generate fuzzy probability maps	Average dice coefficient was 0.87 for their method
El-Hariri et al. [24]	Investigated shadow peak and confidence-weighted structured phase symmetry methods along with investigating U-net and multi-channeled U-net	U-net outperformed image processing methods and had average dice coefficients of 0.86 and 0.92 on different datasets
Baka et al. [25]	Used the U-net and compared to random forest algorithm for segmentations	U-net outperformed the random forest algorithm and had a F-score of 0.90 vs. 0.83 for random forest

Table 1.1 A summary of the literature review of methods of segmentation bone surfaces from US data.

The results of these studies demonstrate that although methods based on deep learning have shown successful results, they are dependent on the amount of training data used. This is specifically an issue when trying to develop methods for processing US data due to the data collection method being manual and user-dependent, along with the dependence on the machine settings and patient characteristics which can affect the appearance of the bone surface in the US data. Methods of improving the success of the deep learning algorithms can be using multi-feature images as input [8, 23], or by collecting a large amount of data. However, collecting large amounts of real *in vivo* bone surface US data is not feasible which is why this thesis investigates the use of using large-scale simulated bone surface US data to improve the deep learning methods.

1.6 Thesis Objective

In order to investigate the problem of limited data for training neural networks for orthopedic surgery use, this works looks into evaluating how training a network on large scale simulated data can improve the accuracy and robustness of bone segmentation from real US data. The objective of the study is to demonstrate that by incorporating simulated bone surface US data in the trained work, the success of traditional deep learning methods increases compared to using limited real *in vivo* US data. This method will use a deep learning-based approach to automatically segment the US data which removes the need of

using image processing methods to get rid of the noise and other imaging artifacts. The proposed work will also investigate the effectiveness of training a deep learning network with a transfer learning method where the network will be pre-trained using simulated US data and then the deeper layers will be optimized using real *in vivo* US data. The overall goal of the study is to show that using large scale simulated US data in conjunction with limited real US data will improve the performance of the network for automatically segmenting bone surface images from *in vivo* US data. Our second goal is to show how the use of transfer learning affects the accuracy of the automatic bone surface segmentations when compared to the traditional deep learning approach from real B-mode US images.

1.8 References

1. The Burden of Musculoskeletal Diseases in the United States (BMUS). (n.d.). Retrieved from https://www.boneandjointburden.org/fourth-edition/usbji

2. Sugano, N. (2003). Computer-assisted orthopedic surgery. In *Journal of Orthopaedic Science*. https://doi.org/10.1007/s10776-002-0623-6

2 Zheng, G., & Nolte, L. P. (2015). Computer-Assisted Orthopedic Surgery: Current State and Future Perspective. *Frontiers in Surgery*. https://doi.org/10.3389/fsurg.2015.00066

3 Hacihaliloglu, I. (2017). Ultrasound imaging and segmentation of bone surfaces: A review. *TECHNOLOGY*, *05*(02), 74–80. https://doi.org/10.1142/s2339547817300049

4 Schumann S. (2016) State of the Art of Ultrasound-Based Registration in Computer Assisted Orthopedic Interventions. In: Zheng G., Li S. (eds) Computational Radiology for Orthopaedic Interventions. Lecture Notes in Computational Vision and Biomechanics, vol 23. Springer, Cham

5 Kowal, J., Amstutz, C., Langlotz, F., Talib, H., & Ballester, M. G. (2007). Automated bone contour detection in ultrasound B-mode images for minimally invasive registration in computer-assisted surgery - An in vitro evaluation. *International Journal of Medical Robotics and Computer Assisted Surgery*, *3*. https://doi.org/10.1002/rcs.160

6 Hacihaliloglu, I., Abugharbieh, R., Hodgson, A. J., & Rohling, R. N. (2009). Bone Surface Localization in Ultrasound Using Image Phase-Based Features. *Ultrasound in Medicine and Biology*. https://doi.org/10.1016/j.ultrasmedbio.2009.04.015

7 Wang, P., Patel, V. M., & Hacihaliloglu, I. (2018). Simultaneous Segmentation and Classification of Bone Surfaces from Ultrasound Using a Multi-feature Guided CNN. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. https://doi.org/10.1007/978-3-030-00937-3_16

8 Salehi, M., Prevost, R., Moctezuma, J. L., Navab, N., & Wein, W. (2017). Precise ultrasound bone registration with learning-based segmentation and speed of sound calibration. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. https://doi.org/10.1007/978-3-319-66185-8_77

9 Raghu M, Zhang C, Kleinberg J, Bengio S. Transfusion: Understanding transfer learning for medical imaging. InAdvances in Neural Information Processing Systems 2019 (pp. 3342-3352).

10 Frid-Adar, M., Diamant, I., Klang, E., Amitai, M., Goldberger, J., & Greenspan, H. (2018). GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. *Neurocomputing*. https://doi.org/10.1016/j.neucom.2018.09.013

11 Zhao, A., Balakrishnan, G., Durand, F., Guttag, J. v., & Dalca, A. v. (2019). Data augmentation using learned transformations for one-shot medical image segmentation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. https://doi.org/10.1109/CVPR.2019.00874

12 Liu, S., Wang, Y., Yang, X., Lei, B., Liu, L., Li, S. X., Ni, D., & Wang, T. (2019). Deep Learning in Medical Ultrasound Analysis: A Review. In *Engineering*. https://doi.org/10.1016/j.eng.2018.11.020

13 Pandey, P. (2018). *Real-time ultrasound bone segmentation and robust US-CT registration for surgical navigation of pelvic fractures* (Issue December). https://doi.org/10.14288/1.0375839

14 Suhm, N., Jacob, A. L., Nolte, L. P., Regazzoni, P., & Messmer, P. (2000). Surgical navigation based on fluoroscopy - Clinical application for computer-assisted distal locking of intramedullary implants. *Computer Aided Surgery*. https://doi.org/10.1002/igs.1001

15 Hacihaliloglu, I. (2010). *Towards A Novel Minimally Invasive 3D Ultrasound Imaging Based Computer Assisted Orthopaedic Surgery System for Bone Fracture Reduction* (Issue April).

16 Gausden, E. B., Christ, A. B., Zeldin, R., Lane, J. M., & McCarthy, M. M. (2017). Tracking Cumulative Radiation Exposure in Orthopaedic Surgeons and Residents. *Journal of Bone and Joint Surgery - American Volume*. https://doi.org/10.2106/JBJS.16.01557

17 C.E. Blane, J.S. Desmond, M.A. Helvie, B.J. Zink, J.E. Bailey, L.D. Yang, N.R. Dunnick, "Academic radiology and the emergency department: does it need changing?", Academic Radiology, vol.14, no.5 pp. 625-630, 2007.

18 Klibanov, A. L., & Hossack, J. A. (2015). Ultrasound in radiology: From anatomic, functional, molecular imaging to drug delivery and image-guided therapy. In *Investigative Radiology*. https://doi.org/10.1097/RLI.0000000000000188

19 Daanen, V., Tonetti, J., & Troccaz, J. (2004). A fully automated method for the delineation of osseous interface in ultrasound images. *Lecture Notes in Computer Science*, 549–557. https://doi.org/10.1007/978-3-540-30135-6_67

20 Jia, R., Mellon, S. J., Hansjee, S., Monk, A. P., Murray, D. W., & Noble, J. A. (2016). Automatic bone segmentation in ultrasound images using local phase features and dynamic programming. *Proceedings - International Symposium on Biomedical Imaging*. https://doi.org/10.1109/ISBI.2016.7493435

21 Villa, M., Dardenne, G., Nasan, M., Letissier, H., Hamitouche, C., & Stindel, E. (2018). FCN-based approach for the automatic segmentation of bone surfaces in ultrasound images. *International Journal of Computer Assisted Radiology and Surgery*, *13*(11), 1707–1716. https://doi.org/10.1007/s11548-018-1856-x

22 Alsinan, A. Z., Patel, V. M., & Hacihaliloglu, I. (2019). Automatic segmentation of bone surfaces from ultrasound using a filter-layer-guided CNN. *International Journal of Computer Assisted Radiology and Surgery*, *14*(5), 775–783. https://doi.org/10.1007/s11548-019-01934-0

23 El-Hariri, H., Mulpuri, K., Hodgson, A., & Garbi, R. (2019). Comparative Evaluation of Hand-Engineered and Deep-Learned Features for Neonatal Hip Bone Segmentation in Ultrasound. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. https://doi.org/10.1007/978-3-030-32245-8_2

24 Baka, N., Leenstra, S., & van Walsum, T. (2017). Ultrasound Aided Vertebral Level Localization for Lumbar Surgery. *IEEE Transactions on Medical Imaging*. https://doi.org/10.1109/TMI.2017.2738612

CHAPTER 2

METHODS

2.1 Overview

The main aim of this thesis is to show how using large-scale simulated US data can improve the performance of bone segmentation. As stated in Chapter 1, this work explores the idea of using simulated US data in order to improve the accuracy and robustness of an already developed neural network called the U-net [1]. One of the ways that simulated US data will be investigated to improve traditional deep learning methods is by evaluating how well a transfer learning method works on the network architecture and if it improves the segmentation output when compared to using limited real US data. The second way that simulated US data will be investigated is by evaluating how well a network trained with a dataset comprised of both large-scale simulated US data and limited real US data performs automatic bone surface segmentation on real US data.

2.2 Data Acquisition

2.2.1 Overview

In order to determine whether transfer learning has an impact on the outcome of a neural network, different networks were trained with two types of datasets. One dataset consisted entirely of real B-mode US images obtained from patients while the second dataset consisted of simulated US data that was obtained via 3D Slicer and the Public software Library for Ultrasound (PLUS) toolkit [2, 3].

2.2.2 *In vivo* B-mode US Images and Manual Segmentations

380 *in vivo* US scans of radius and femur bone surfaces were collected from 5 different subjects after obtaining approval from the Rutgers University Institutional Review Board (IRB). The US scans were then manually segmented by an expert. In order to keep training data separate from the testing data, Subject 5's US scans were used for the testing data while Subjects 1-4's US scans were used as a part of the training dataset. However, in order to check the cross validation of the mixed network, different subjects needed to be used for training and testing data which is described in Section 2.6. To have enough images for training and testing the network, data augmentation was used on all the US images and their corresponding manual segmentations from Subjects 1-5. The data augmentation steps consisted of rotation ±10 degrees and translation in the x and y directions ±5 pixels. These data augmentation steps resulted in a total of 2,000 images across all 5 subjects. 1,600 images were from selected from 4 of the subjects and 400 images from remaining subject were used for the training and testing datasets, respectively. Therefore, the network training was split into an 80:20 percent ratio of training and testing data.

2.2.3 Simulated US Images and Manual Segmentations

3D Slicer, SlicerIGT and PLUS were used in order to obtain simulated US images and their segmentations [4]. STL files of femur and radius bones were loaded in the US simulator in 3D Slicer in order to obtain the simulated US scans [5].

PLUS is a toolkit platform that allows for communication with 3D Slicer in real time by obtaining the US image and sending it as configuration files to 3D Slicer [3]. The configuration files in PLUS contain different acoustic material properties and surface

meshes so that positions of the objects can be tracked during the US simulation [3]. In order to have the correct acoustic properties required for the US, the attenuation and reflection coefficients and speed of sound data in addition to other parameters were also sent over in the configuration file. The US transducer can also be selected and in this study, a linear US transducer, Ultrasonix L9-4/38 was used for the simulation. The machine settings of the transducer were kept within the same ranges as the transducer used for real image acquisition.

3D Slicer was then used to visualize the simulated US images and to perform the manual segmentations (***Fig. 2.1***). The images from the PLUS platform were sent over to Slicer in real-time in order to track the positions of the transducer relative to the bone surface in the bone model STL file. The relative positions were obtained using the Transforms Module in Slicer which allowed the user to control the transducer and move it along the bone surface to obtain simulated US images in real-time. As the transducer was moved along the bone surface, a sequence of simulated US images with their relative spatial positions were generated as 3D volumetric US images. These sequences were able to be saved using the Sequence Browser in 3D Slicer. After obtaining this sequence, the simulated US could then be segmented using the Single Slice Segmentation module and the Segment Editor module in Slicer. The Segment Editor module allowed the user to paint over the bone surface, which effectively created a manual segmentation of the bone surface [6]. The corresponding simulated US image and the segmentation were then able to be exported as PNG files.

a) b)

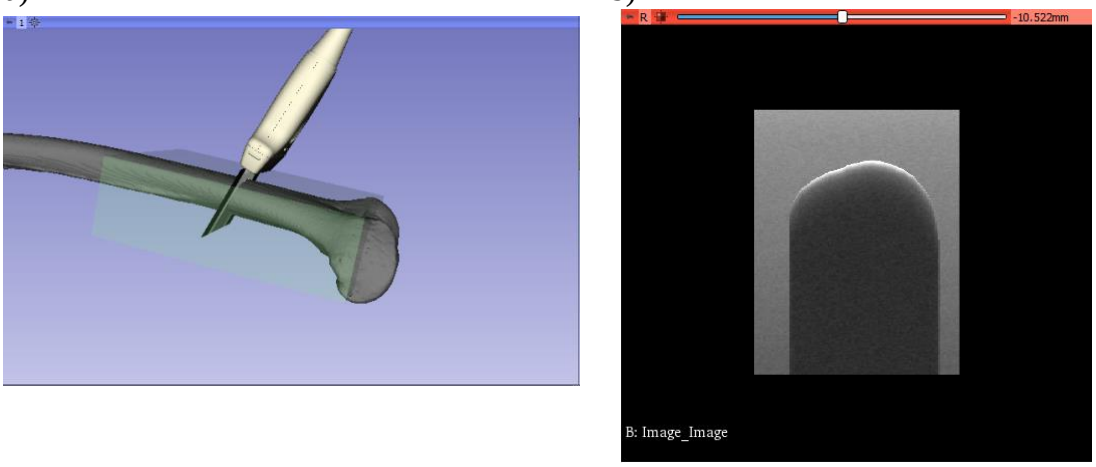


Fig. 2.1 The image on the left shows the US transducer in 3D Slicer along with the femur model and the gel block which allows for the simulation of the US. The image on the right shows the corresponding Simulated US image that is generated in 3D Slicer using PLUS.

3D Slicer and PLUS were used to obtain 355 US images of the femur and radius bone. 167 of the images were from the femur STL file and 188 of the images were from the radius STL file. The femur and radius images from Slicer along with real US images are shown in ***Fig. 2.2***. Out of the total 355 images, 142 images and their manual segmentations were separated in the testing dataset. Data augmentation was also performed on these scans in order to create a total of 10,000 simulated US images and segmentations. The data augmentation steps consisted of the same rotation ±10 degrees and translation in the x and y directions ±5 pixels as the real US dataset. The data was split according to the same 80:20 percent ratio of training and testing data that was used for the real US data

acquisition. Therefore, 8,000 simulated US images were used for training and validation while 2,000 of the simulated US images were used for testing.

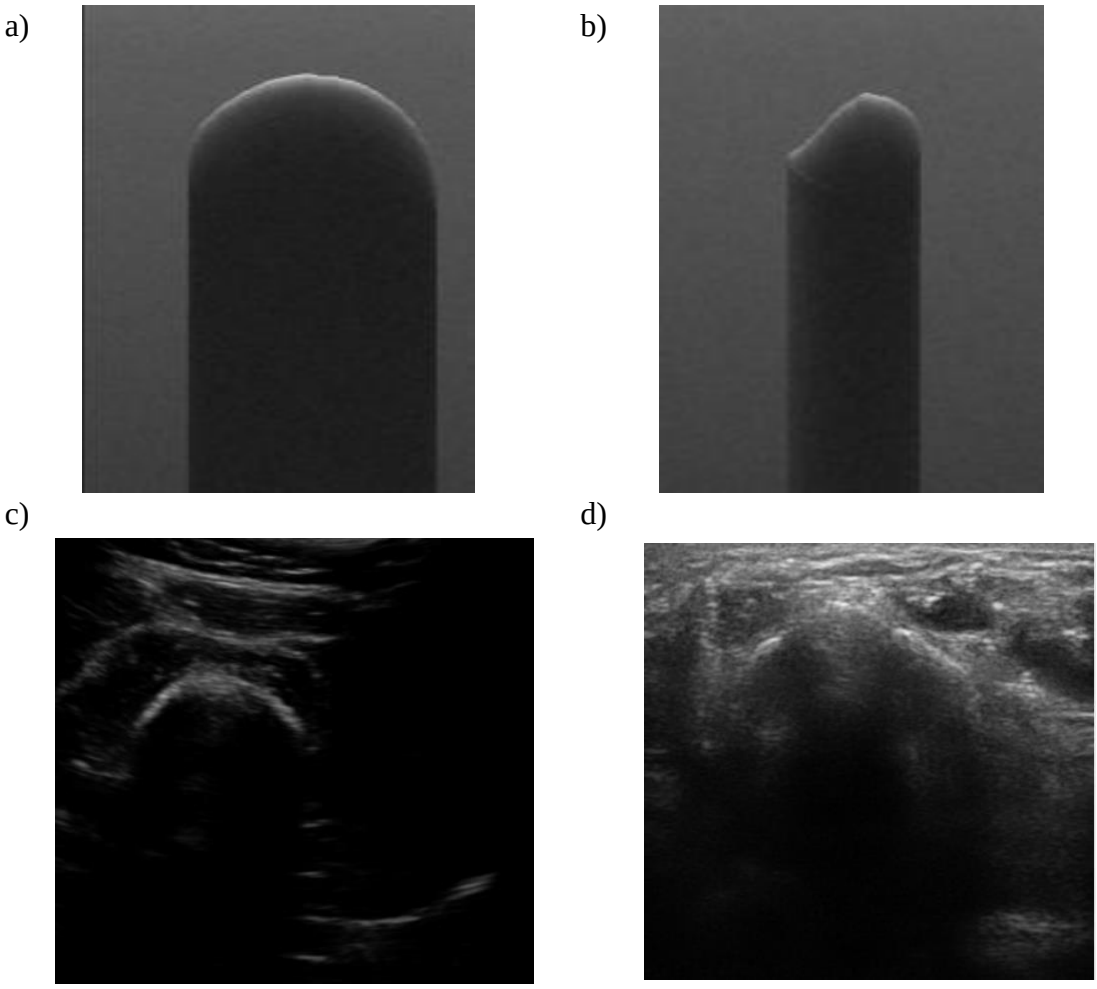


Fig. 2.2 The images above show the simulated US images generated through 3D Slicer and Plus and the B-mode US taken *in vivo.* The image shows a) Simulated Femur US, b) Simulated Radius US, c) Real Femur B-mode US, d) Real Radius B-mode US. The Simulated US images lack the artifacts and noise that is present in the real US images.

2.3 U-net for bone US data segmentation

The U-net neural network was developed for the purpose of making it easier and more efficient to segment biomedical images [1]. The U-net is a type of modified fully convolutional neural network (FCN) that was developed for the segmentation of biomedical images by predicting each pixel's class [1]. The U-net was proposed in order to solve the problem of the requirement of large datasets for training and large networks [1].

Compared to the fully convolutional neural network, the U-net has modified network architecture so that the network is symmetric and consists of 3 separate areas: the downsampling (contracting) path, the bottleneck and the upsampling (expansive) path (***Fig. 2.3***) [7]. The purpose of the downsampling path is to get the context of the input image that is needed to start the segmentation. Therefore, the downsampling path helps to obtain what information is present in the image. The bottleneck is built from 2 convolutional layers and then the upsampling path works to localize the segmentation with the contextual information [7]. The network architecture shows that in the contracting path, each block takes an input, then applies two 3x3 convolution layers and a 2x2 max pooling operation [8]. Convolutional layers use filters to detect patterns in the image and output a feature map. The spatial pooling operations reduce the dimensions of the feature maps while retaining the data but minimizing the number of parameters in the network. The contraction path of the neural network is the basis of any fully convolutional neural network. However, the difference with the U-net is the expansive path. The blocks in the expansive path take the input, apply two 3x3 convolution layers and then apply a 2x2 upsampling layer which increases the

dimensions of the input [8]. The symmetry in the U-net allows for the features in the images to be learned properly so that the same features can be used to reconstruct the image and perform a segmentation [8]. The loss function in the U-net is also different compared to other fully convolutional neural networks since it uses a pixel-based loss weight function. Higher weights are assigned at the border of the segmented surfaces and the loss function works to ensure that the pixels are properly classified into either the segmentation or the background [8].

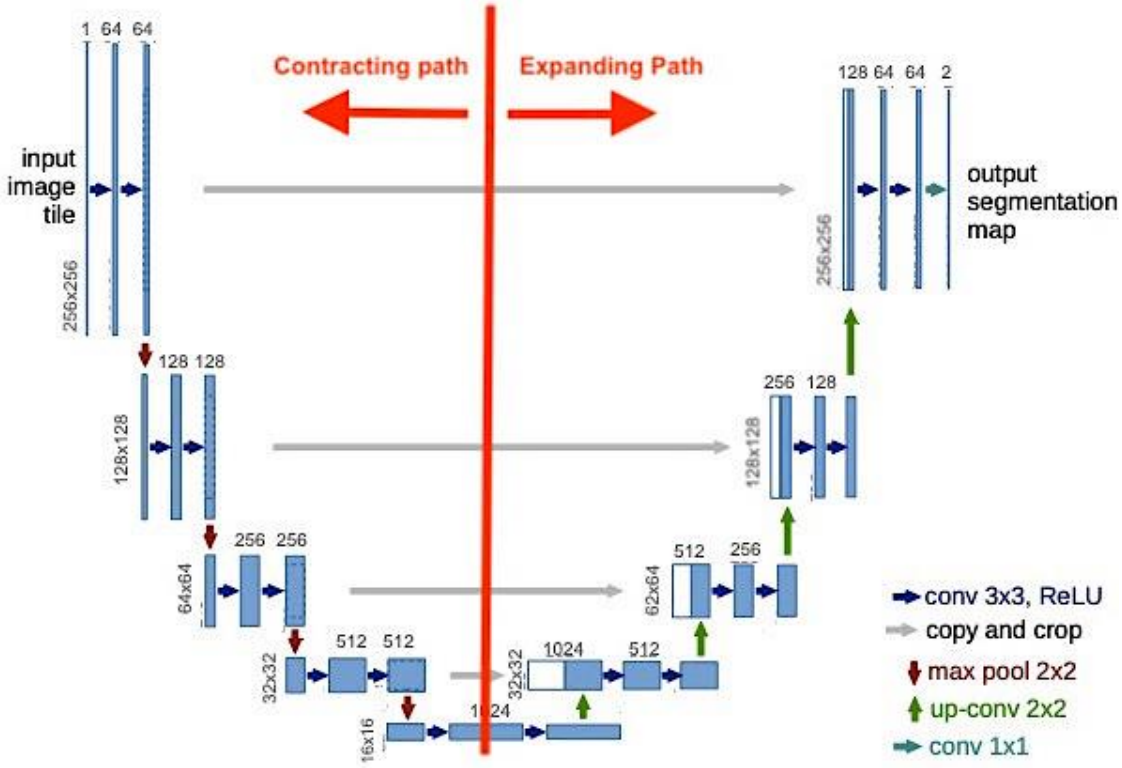


Fig. 2.3 The U-net architecture is shown above. The network consists of a contracting path and expansive path which work together to obtain contextual information from the image and then localize it to perform a segmentation [1,7]

The data that was used for training was separated into different datastores: Real US images, Real US manual segmentations, Simulated US images, and Simulated US manual segmentations. The testing data was also separated into datastores according to the type of US scan – simulated or real US data. As the manual segmentations were binary images, the label IDs for the images were 1 and 0 where 1 denoted the bone surface values and 0 was the background which resulted in only 2 classes for the network. The size of the input images was modified to be 256x256 and the encoder depth was 4. Initially, in order to determine how the training parameters work in MATLAB, different networks were trained using different parameters on only the Real US data. The parameters that were varied for the network included the encoder depth, minibatch size, learning rate, and the number of epochs. The L2 regularization was changed for one network but all other networks kept the default value of 0.0001. The parameters were chosen to be varied depending on how well the trained network performed on the validation data set.

2.4 Transfer Learning

Transfer learning is a part of machine learning algorithms where an algorithm learns information via a "source task" and applies to another related "target task" [9]. Using this method, a neural network would not need to be trained from scratch and can instead use the features that it learned from the source task and apply it to the target task [10]. For the purposes of this research, transfer learning can help optimize a network with minimal real US data available. The neural network can be trained on large scale simulated US data which will allow it to learn general US features. Then, the deeper layers of the network

will be retrained using only limited real US data which will allow the network to segment real US bone surfaces.

In order to investigate the use of using simulated US data for transfer learning methods, a network trained on large-scale simulated US data was optimized by retraining the final convolutional layer, the pixel classification layer, and softmax layers with limited real US data. The deeper layers allow for learning of the features of the images and the final convolutional layer is one of the last layers that is trained therefore, it was chosen to be retrained.

2.5 Validation Metrics

Quantitative evaluation was performed by calculating the Sorensen Dice coefficient and Average Euclidean Distance.

The Sorensen Dice coefficient was automatically calculated in MATLAB using a built-in function [11]. The equation used in the calculation for this coefficient is shown in ***Equation 1***. The Dice Coefficient calculation takes into account the true positives, false positives, and false negatives while comparing the automatic segmentations to the manual segmentations. The true positives account for the amount of overlap between the two types of segmentations while the false positives and false negatives work to lower the dice coefficient value for mistakes in the automatic segmentations.

$$dice(A,B) = \frac{2\times TP}{2\times TP+FP+FN} \qquad (\textbf{\textit{Equation 1}})$$

Equation used in the calculation of the Sorensen Dice Coefficient. A and B in this equation represent the manual segmentation and the automatic segmentation. [11].

The Euclidean distance error is the distance of the line segment separating a point on the manual segmentation from the closest point on the automatic segmentation. The Euclidean distance only gives the information of how far apart the automatic and manual segmentations are however, in order to have a good metric, it needs to incorporate if the entire segmentation is properly covered as well. Therefore, the Euclidean distance error was modified by taking into account the length of the segmentations. The Euclidean distance, which is reported in millimeters (mm), was divided by the ratio of the overlapping length of the pixels in both the manual and automatic segmentation to the sum of the length of the manual and automatic segmentations. In this thesis, the metric is referred to as the Average Euclidean Distance (AED) and it is reported in mm. A higher AED means that the manual and automatic segmentations are far apart in distance or are not the same coverage and a lower AED means that both the segmentations are overlapping. Both the Dice and AED were used to evaluate how well the automatic segmentation matches the manual segmentation. An evaluation of the validation metrics is shown in Chapter 3.8

2.6 K-fold Cross Validation

In order to evaluate if the network was performing well and not over-fitting the data, a method of validation called the K-fold cross validation was used on the mixed network developed with Simulated and Real US Data. Cross validation allows for the resampling of the data in order to evaluate a model [12]. K-fold cross validation is a method of cross validation where different data in the original dataset is allowed to appear in the testing and training dataset [12]. The method depicted in ***Fig. 2.5*** requires that a certain number of iterations (k) are run on the original dataset where randomly, certain data are

used for validation and the remaining for training. The average of the validation results for the different iterations is taken to get an overall metric for how well the network is trained [12].

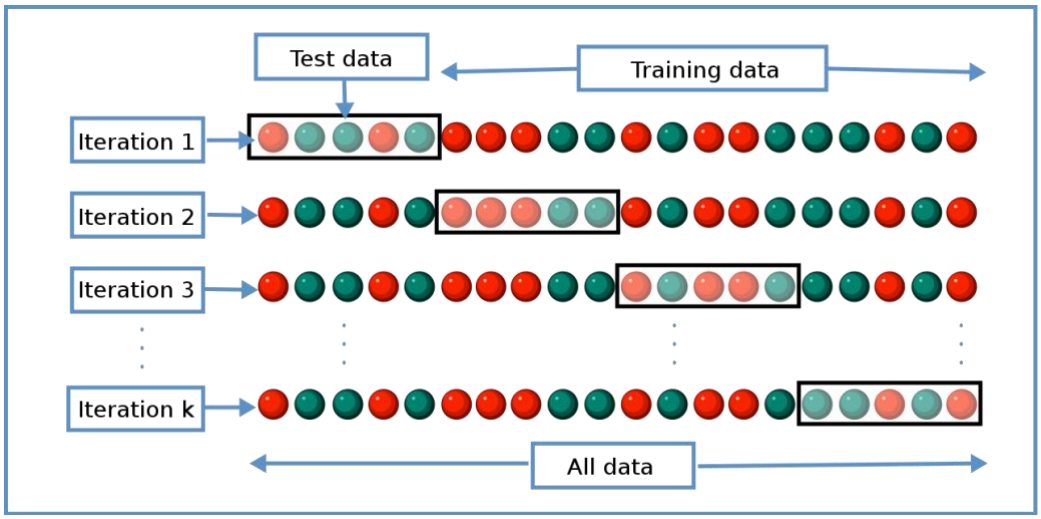


Fig. 2.4 The process of K-fold Cross Validation is depicted in the image above. The different number of iterations (denoted as k) are run on the original dataset and the testing data is randomized per iteration. The average of the results of validation process is then taken to determine the performance of the network [13].

For the purpose of this thesis, a 3-fold cross validation was performed on the mixed network that was trained using both simulated and real US data. Since all of the simulated US data was used as part of training (10,000 images), the real US data was varied. All of the iterations used 1600 real US images for training and 400 real US images for testing. For the first iteration, real US data from Subjects 1-4 was used for training and testing was done on Subject 5's data which was composed of 30 original US scans which were augmented to 400 images. The second iteration used Subjects 1,2,3 and 5 for training and Subject 4 for testing. Subject 4 originally had 35 real US scans. The last iteration used

Subjects 1,2,4 and 5 for training and Subject 3 for testing. Originally, Subject 3 had 70 original real US scans that were then augmented. The 3-fold cross validation performed on the mixed network will demonstrate how well the network was trained and if it is able to segment different bone surface data that the network has not seen in the training dataset.

2.7 Training Networks and Validation Process

2.7.1 Training on Real US Data and Testing on Real US Data

After the Real US data network was trained on the training dataset, validation was performed on the same 1,600 images from Subjects 1-4 by checking the Sorensen Dice coefficient between the automatic segmentations and the manual segmentations. In order to evaluate how the network performed on the testing data, the Sorensen Dice coefficient and Distance index were calculated on the automatic segmentation of the 400 testing images and their corresponding manual segmentations.

2.7.2 Training on Simulated US Data and Testing on Simulated US Data

After the parameter optimization was performed on the network trained on Real US data, the same parameters were used to train a new U-net neural network on the simulated US data. The neural network was also validated on the 8,000 training images of the Simulated US data and their corresponding segmentations and their Dice and Distance index were calculated. The network was then evaluated against the 2,000 testing images of the Simulated US data and the Distance index and Sorensen Dice Coefficient were calculated.

2.7.3 Training on Simulated US Data and Testing on Real US Data

The network that was trained on the simulated US data was then also tested on the Real US images. 400 of the testing images from Subject 5 were used to perform the evaluation. The Distance index and Sorensen Dice Coefficients were calculated in order to determine how well a network that is trained on Simulated US data perform on the real US data.

2.7.4 Training on Transfer Learning Network and Testing on Real US Data

Another network that was trained was the transfer learning network which used a base of the network trained on simulated US data. Using the network that was trained on the Simulated US data, the final layers of the network, such as the final convolutional layer, were retrained with 260 images from the Real US training dataset (specifically, from Subject 1). The network was then tested on the 400 testing images of the Real US dataset which were from Subject 5. The Dice coefficient and Distance index values were calculated for that dataset as well in order to compare the automatic segmentations to their respective manual segmentations.

2.7.5 Training on Mixed Network and Testing on Real US Data

In order to evaluate how well the large-scale simulated US data and limited real US data worked to segment the real US data, a network was trained from scratch using both types of data. All 10,000 images from the simulated US data and 80% of the real US data (1,600) were used to train a neural network using similar parameters as the previous networks. This mixed network was validated against the 400 images (20%) of real US data.

2.8 References

1. Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. https://doi.org/10.1007/978-3-319-24574-4_28

2. Kikinis, R., Pieper, S. D., & Vosburgh, K. G. (2014). 3D Slicer: A Platform for Subject-Specific Image Analysis, Visualization, and Clinical Support. In *Intraoperative Imaging and Image-Guided Therapy*. https://doi.org/10.1007/978-1-4614-7657-3_19

3. Lasso, A., Heffter, T., Rankin, A., Pinter, C., Ungi, T., & Fichtinger, G. (2014). PLUS: Open-source toolkit for ultrasound-guided intervention systems. *IEEE Transactions on Biomedical Engineering*. https://doi.org/10.1109/TBME.2014.2322864

4. Ungi, T., Lasso, A., & Fichtinger, G. (2016). Open-source platforms for navigated image-guided interventions. In *Medical Image Analysis*. https://doi.org/10.1016/j.media.2016.06.011

5. Bartha, L., Lasso, A., Pinter, C., Ungi, T., Keri, Z., & Fichtinger, G. (2013). Open-source surface mesh-based ultrasound-guided spinal intervention simulator. *International Journal of Computer Assisted Radiology and Surgery*. https://doi.org/10.1007/s11548-013-0901-z

6. Segment editor. (2017). Retrieved from https://slicer.readthedocs.io/en/latest/user_guide/module_segmenteditor.html

7. U-Net. (2018, June 15). Retrieved from http://deeplearning.net/tutorial/unet.html

8. Sankesara, H. (2019, January 23). U-Net. Retrieved from http://towardsdatascience.com/u-net-b229b32b4a71

9. Torrey, L., & Shavlik, J. (2009). Transfer learning. In *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*. https://doi.org/10.4018/978-1-60566-766-9.ch011

10. Brownlee, Jason. “Transfer Learning in Keras with Computer Vision Models.” *Machine Learning Mastery*, 15 May 2019, machinelearningmastery.com/how-to-use-transfer-learning-when-developing-convolutional-neural-network-models/.

11. dice. (n.d.). Retrieved from https://www.mathworks.com/help/images/ref/dice.html

12. Sanjay.M. (2018, November 13). Why and how to Cross Validate a Model? Retrieved from https://towardsdatascience.com/why-and-how-to-cross-validate-a-model-d6424b45261f

13. File:K-fold cross validation EN.svg. (n.d.). Retrieved from https://commons.wikimedia.org/wiki/File:K-fold_cross_validation_EN.svg

CHAPTER 3

RESULTS AND DISCUSSION

3.1 Overview

In this chapter, the results of the evaluation of the different networks will be shown both quantitively and qualitatively. The networks' evaluation was performed by using the validation metrics discussed in Chapter 2.5. The different validation metrics will be compared to determine the effect that transfer learning and mixed network had on the performance of a neural network compared to networks trained solely on one type of US data. The results will help determine what method of network optimization works best to automatically segment the bone surfaces in US data and whether simulated data can be used for the segmentation of real US bone surface data. Additionally, this section will answer the questions discussed at the end of Chapter 1.1:

1) Can simulated US data be used to train a neural network for accurate and robust segmentation of real i*n vivo* US data?
2) Can simulated and real *in vivo* US data be used together to train a neural network for accurate and robust segmentation of real in vivo US data?

3.2 Evaluation of Real US Data Based Network

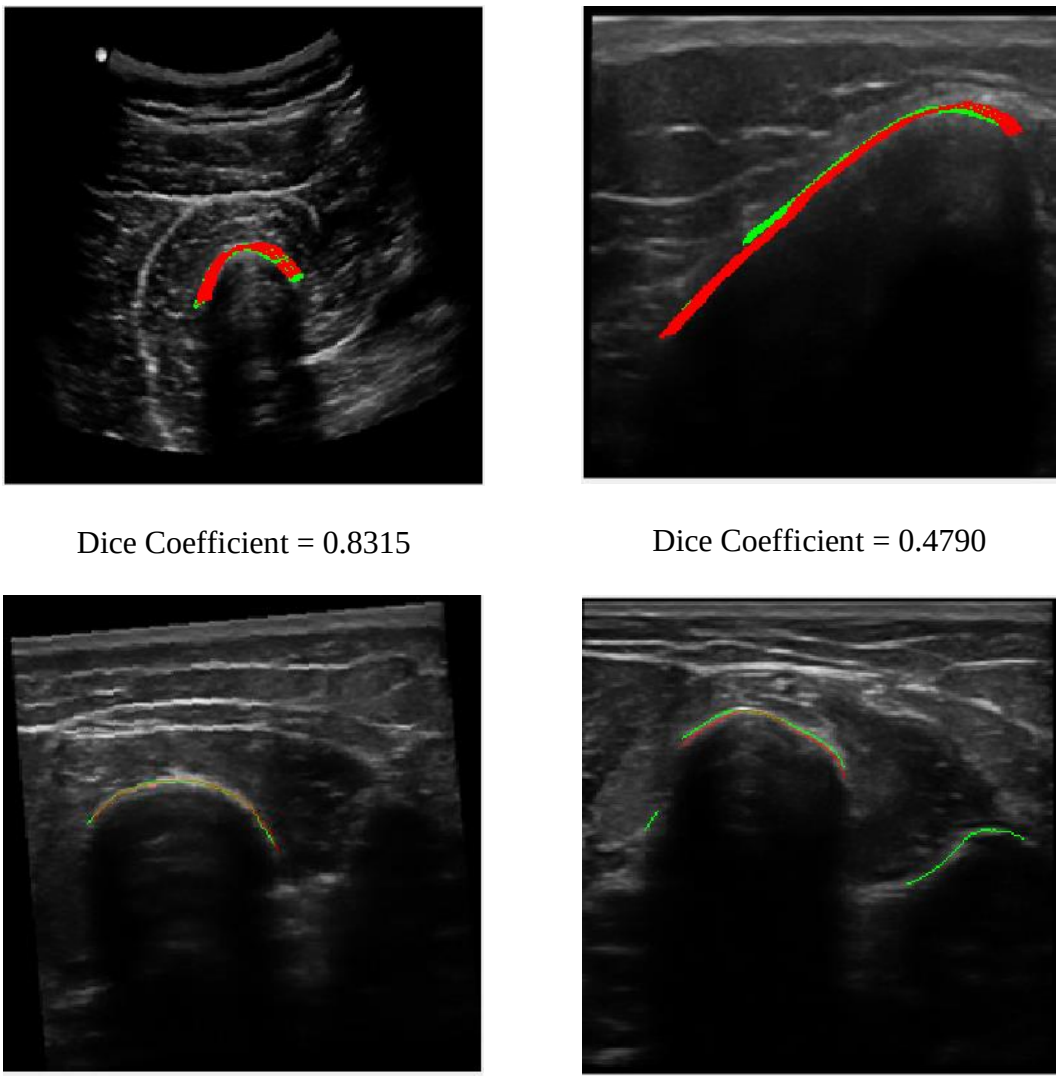
The first step to the development of a neural network for segmenting US bone surfaces involved experimenting with the different parameters that are used in the optimization of the network. This step involved training different networks using real US data to determine what the optimal parameters for developing future networks would be. ***Table 3.1*** shows a few examples of the different networks that were trained on real US data

and their validation and testing results. From the network optimization, trainedNet_1600T11 was selected as the optimized network because of the high dice coefficient on testing data and was then chosen to be used for comparison to the Simulated US network, transfer learning network and the mixed data network.

Network Name	Parameters Varied	Dice Coefficient on Training Data	Dice Coefficient on Testing Data
trainedNet_1600T6	Epochs = 50 Learn Rate = 3e-4	0.78883	0.6522
trainedNet_1600T8	Epochs = 80 Learn Rate = 3e-4	0.8659	0.657
trainedNet_1600T10	Epochs = 30 Learn Rate = 6e-4	0.7334	0.6168
trainedNet_1600T11	Epochs = 100 Learn Rate = 3e-4	0.7643	0.7067

Table 3.1 shows the different networks that were trained on real US data. In order to optimize the network, the learning rate and number of epochs were varied during training. TrainedNet_1600T11 had the best dice coefficient on the testing data and was chosen for comparison against the other networks.

The quantitative evaluation of the network was done by determining the best and worst dice coefficient and AED values. The corresponding automatic segmentation images were there overlaid on the original US image and the corresponding manual segmentation. ***Fig. 3.1*** shows the images with the best and worst Dice coefficients and AED values.



Dice Coefficient = 0.8315 Dice Coefficient = 0.4790

AED = 0.0130 mm AED = 8.2942 mm

Fig. 3.1 The top row shows the segmentations on real US Data with the maximum (left) and minimum (right) dice coefficient value. The bottom row shows the minimum (left) and maximum (right) AED values. The green line is the automatic segmentation generated by the network and the red line is the manual segmentation.

The quantitative images in ***Fig. 3.1*** show that only using limited real US data to train a network does not give optimal segmentations. As seen by the images with the

minimum values of the dice coefficient and maximum AED values, the optimized algorithm does not segment the bone surface accurately. In ***Fig. 3.1***, the automatic segmentation in the minimum dice coefficient image missed a large portion of the bone surface towards the bottom of the image. Additionally, in ***Fig. 3.1***, the automatic segmentation in the maximum AED image chose part of the soft tissue interface as the bone surface. These mistakes in segmentations would result in an improper segmentation result which can be detrimental to the use of the scans in surgery.

3.3 Evaluation of the Simulated US Data Based Network

3.3.1 Simulated US Data Based Network Tested on Simulated US Data

After a network was trained on real US data, simulated US data was used to train a network keeping similar parameters as the optimized real US data-based network. This optimized network trained on simulated US data was then tested on the 20% of the testing dataset of the simulated US data. ***Fig. 3.2*** shows the images for the best and worst dice coefficient and AED segmentations from the simulated data.

Overall, the network that was trained on simulated US data and tested on simulated US data performed very well. In ***Fig. 3.2,*** even the automatic segmentation for the image with the minimum dice coefficient value was good and it could be used for surgical purposes. However, the result of a dice coefficient of 0.5865 does not sufficiently summarize how well the actual automatic segmentation was because according to the image, the segmentation lines up very well with the bone surface and the manual segmentation.

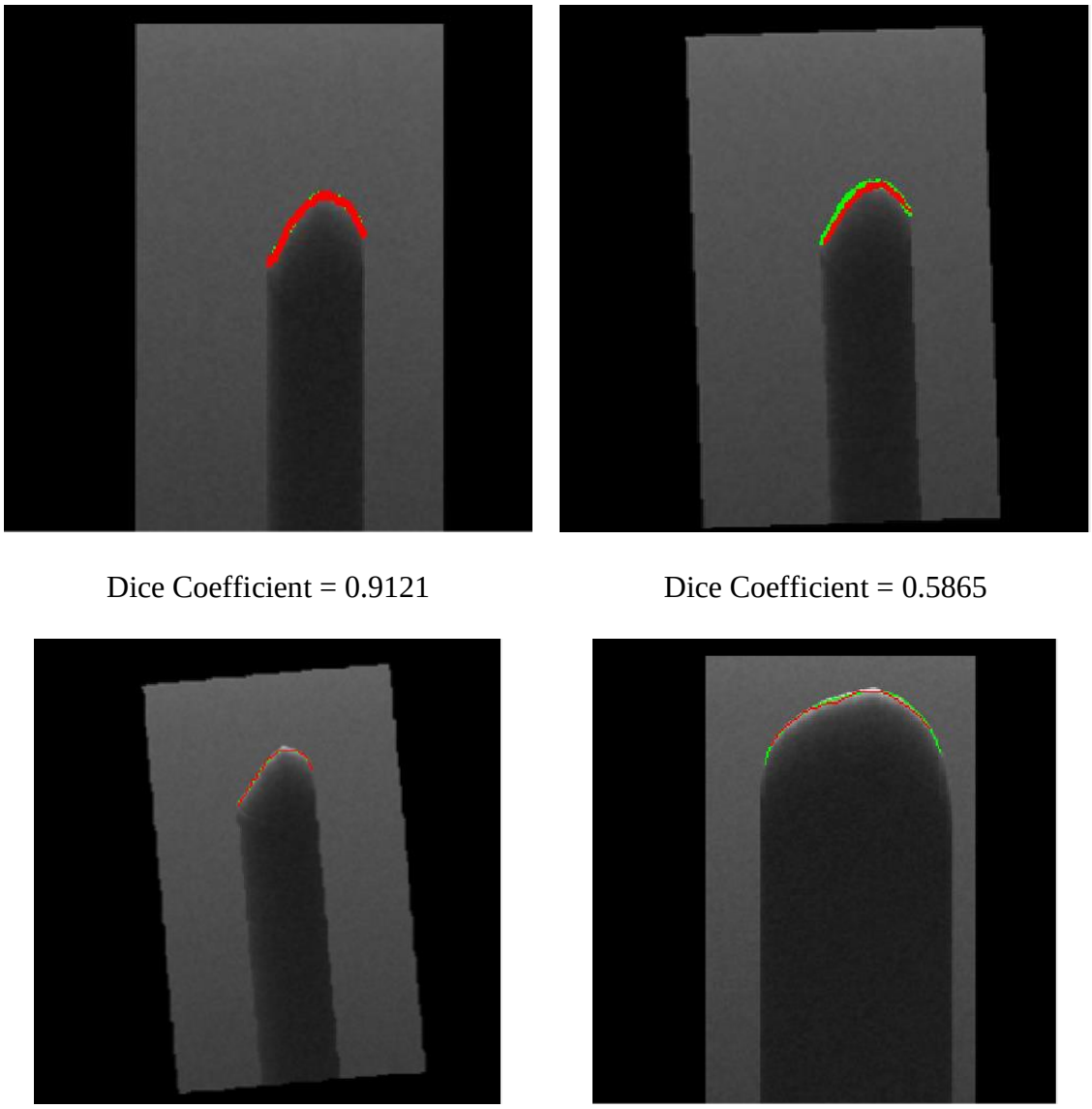
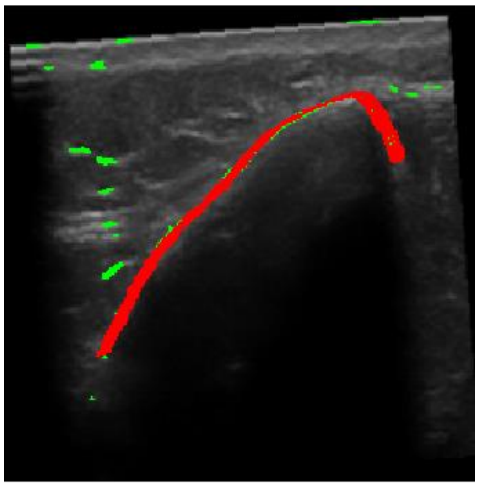


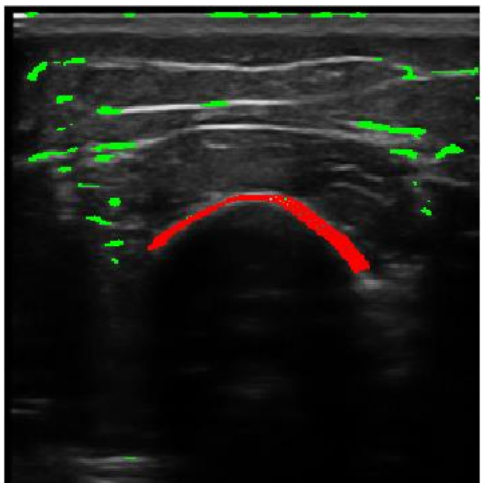
Fig. 3.2 The top row shows the segmentations on simulated US data with the maximum (left) and minimum (right) dice coefficient value. The bottom row shows the minimum (left) and maximum (right) AED values. The green line is the automatic segmentation generated by the network and the red line is the manual segmentation.

3.3.2 Simulated US Data Based Network Tested on Real US Data

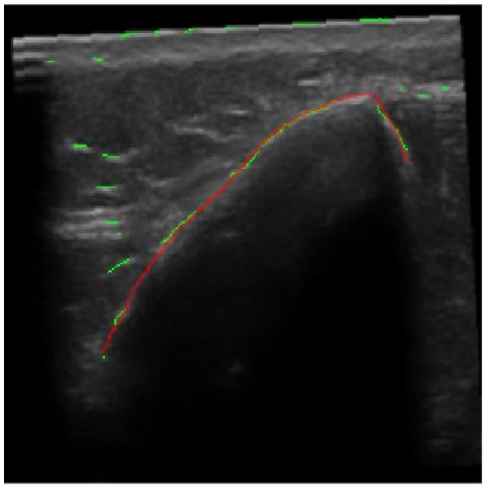
The same optimized network trained on simulated US data was tested on the real US data in order to evaluate how well a network trained solely on large-scale simulated performs on limited real US data. The qualitative results of the dice coefficient and AED are summarized in ***Fig. 3.3***.



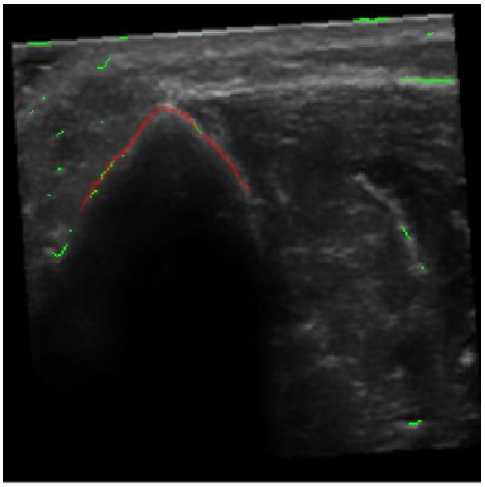
Dice Coefficient = 0.4265



Dice Coefficient = 0.0392



AED = 5.6022 mm



AED = 66.3558 mm

Fig. 3.3 The top row of the figure shows the segmentations on real US data with the maximum (left) and minimum (right) dice coefficient value. The bottom images show the segmentations with the minimum (left) and maximum (right) AED values. The green line is the automatic segmentation generated by the network and the red line is the manual segmentation. This network was trained on simulated US data and therefore, the values are much lower compared to the other networks.

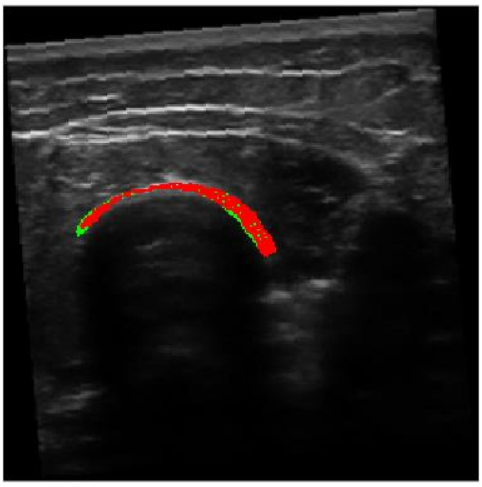
Overall, these three networks (network trained on real US data and tested on real data, network trained on simulated US data and tested on both simulated data and real data) show how well the U-net performs on segmenting bone surfaces in US data. A network trained and tested on its own type of dataset performs much better than a network that is trained on simulated US data and tested on real US data (as seen in ***Fig. 3.3***). The discrepancy is understood because the simulated US images do not have the artifacts that are present in the real US images causing the algorithm to fail at segmenting the specific bone surface. The failure in segmenting the actual bone surface results in very low dice and high AED values.

The next two sections will show the results of a network trained via transfer learning and network trained with both types of data.

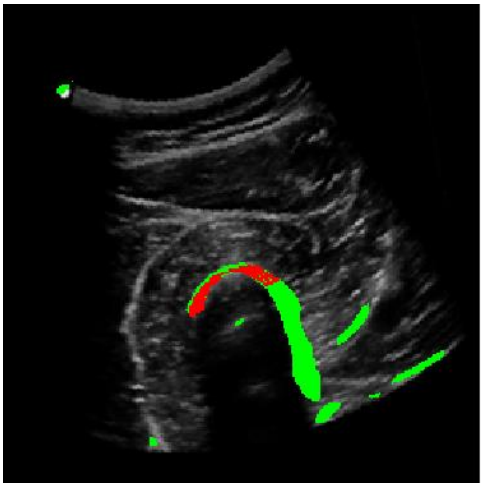
3.4 Evaluation of the Transfer Learning Network

The transfer learning network is a network that has the base of the simulated US network with the last layers of the network retrained on the real US data. Therefore, the network has been taught how to segment both the simulated US data and limited real US

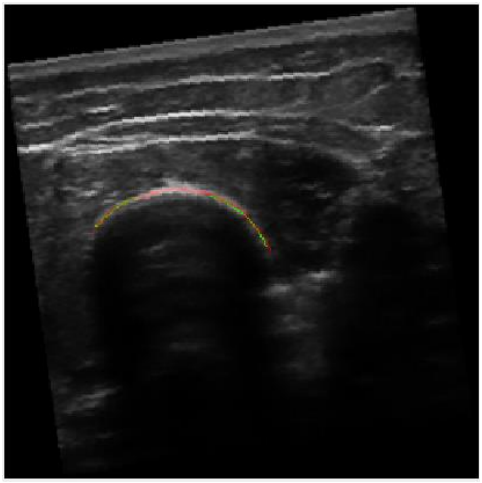
data. However, since the simulated data is large-scale and consists of 8000 scans the results on the real US network were not optimal as seen by the images in ***Fig. 3.4***.



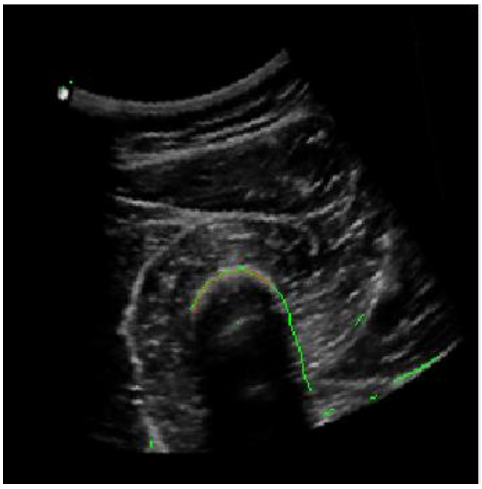
Dice Coefficient = 0.7806



Dice Coefficient = 0.2750



AED = 0.0091 mm



AED = 14.7680 mm

Fig. 3.4 The images above show the segmentations generated by the transfer learning network on real US data. The maximum (left) and minimum (right) dice coefficient value images are show in the top row and the minimum (left) and maximum (right) AED values are shown in the bottom row. The green line is the automatic segmentation generated by the network and the red line is the manual segmentation. The images show that the

algorithm was not able to learn which areas were the bone surface and which showed soft tissue as seen by both of the images with the minimum values.

The purpose of using transfer learning was to evaluate how well using large-scale simulated US data and limited real US data can work together using the deeper layers of the network. The images shown in ***Fig. 3.4*** were generated by using the same network that was trained completely on the large-scale simulated US data and then retraining was done on the final convolutional layer and the pixel classification layer using the real US data. However, the resulting automatic segmentations were not exact since the network missed some of the bone surfaces and segmented other surfaces as well as seen in ***Fig. 3.4***.

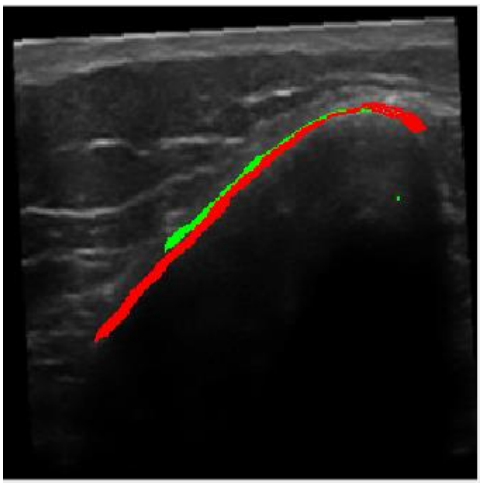
3.5 Evaluation of the Mixed Network

The mixed network is a network that is trained on scratch from 10,000 of the scans from simulated US data and 80% of the total dataset for the real US data. Therefore, this network has exposure from both datasets in all the layers of the network. The qualitative results of the network tested on the 20% of the real US data are shown in ***Fig. 3.5***.

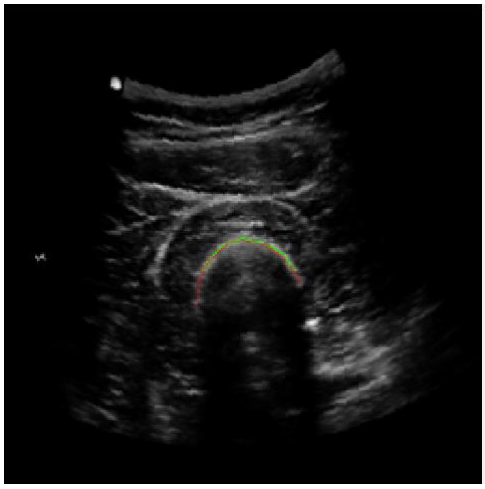
The images in ***Fig. 3.5*** show that at best, the network performs on par with the network trained only on simulated US data and tested on simulated US data but the algorithm still classifies some of the soft tissue surface as the bone surface. The mixed network outperformed the network trained solely on the limited real US data based off of the AED values. The minimum dice coefficient value for the mixed network was 0.2957 compared to a value of 0.4790 which was the minimum dice coefficient for the network

trained from real US data. However, as determined before, the dice coefficient metric might not be the most suitable metric for the thin bone surfaces.

Dice Coefficient = 0.8471



Dice Coefficient = 0.2957



AED = 0.0030 mm

AED = 6.3988 mm

Fig. 3.5 The images above show the segmentations generated by the mixed network which consists of both real and simulated US data which was tested on real US data. The top row shows the segmentations with the maximum (left) and minimum (right) Dice Coefficient value. The bottom row shows the segmentations with the minimum (left)

and maximum (right) AED index values. The green line is the automatic segmentation generated by the network and the red line is the manual segmentation.

3.6 K-Fold Cross Validation of the Mixed Network

Network Name	Testing Dice Coefficient	Testing AED Value (mm)	Notes
MixedNetwork1	0.676 ± 0.114	0.444 ± 0.886	Tested on Subject 5; 30 original US scans
MixedNetwork2	0.771 ± 0.063	0.423 ± 1.083	Tested on Subject 4; 35 original US scans
MixedNetwork3	0.688 ± 0.117	0.102 ± 0.344	Tested on Subject 3; 70 original US scans
Average	0.712 ± 0.052	0.323 ± 0.040	-

Table 3.2 shows the results of the K-fold cross validation on the mixed networks. The networks were all tested on different subjects from the real US dataset.

The results of the K-fold cross validation are summarized in ***Table 3.2***. All of the networks were trained from scratch using all 10,000 images of the simulated US data and 80% (1,600 images) of the real US data. The networks were tested on different Subjects from the real US dataset in order to get an overview of how well the network was performing. The average dice coefficient of the different networks was 0.712 and the average AED value was 0.323 which showed that the networks were automatically segmenting the US bone surfaces efficiently no matter which subject was used in the testing and training datasets.

3.7 Quantitative Comparison of all Networks

Network Name and Type	Average Dice Coefficient on Testing Data	Average AED on Testing Data (mm)
Real US Network Tested on Real US Data	0.707 ± 0.084	1.343 ± 2.001
Simulated US Network Tested on Simulated US Data	0.795 ± 0.050	0.112 ± 0.033
Simulated US Network Tested on Real US Data	0.158 ± 0.070	18.233 ± 8.279
Transfer Learning Network	0.586 ± 0.111	3.060 ± 3.309
Mixed Network	0.676 ± 0.114	0.444 ± 0.886

Table 3.3 shows the results on the testing results for the different networks that were trained. The mixed network showed better results than the limited real US data network according to the AED.

Comparing the testing results from all the different networks, as seen in ***Table 3.3***, the network that performed the best was the network that was trained on simulated US data and tested on simulated US data. This result was expected as the simulated US data does not contain any artifacts that are found in the real US data which makes the automatic segmentation a much more efficient process. In line with this observation is also the fact that the worst performing network was the network that was trained only on simulated US data and was tested on the real US data.

The next two graphs demonstrate the individual validation metrics for all the different networks that were trained.

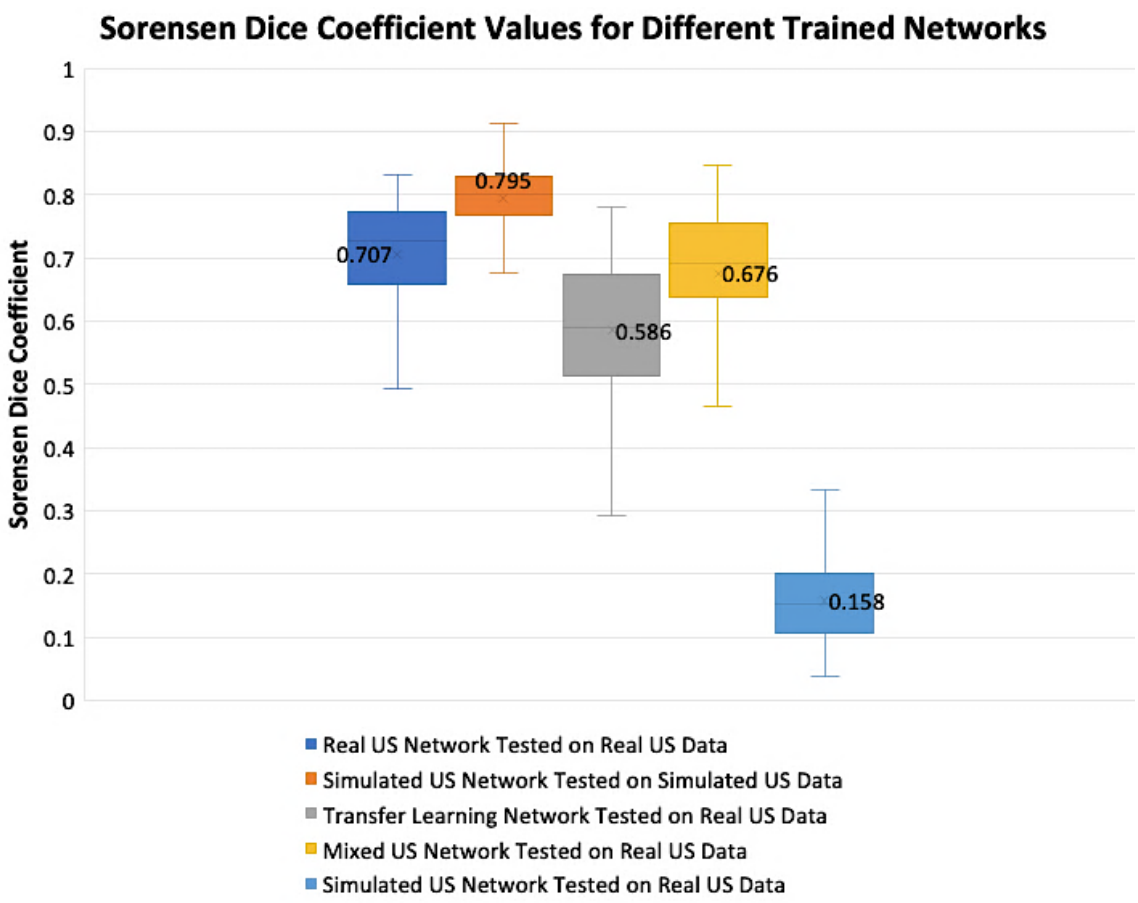


Fig. 3.6 This box and whisker plot shows the dice coefficient values for all of the different trained networks. The lowest average dice coefficient value comes from the network trained on simulated US data and tested on real US data. The mixed network outperformed the transfer learning network in terms of average dice coefficient and the average value is close to the average dice coefficient for the real US network.

The box and whisker plot in ***Fig. 3.6*** shows the dice coefficient values for the different networks that were trained. The best performing network was the network trained on simulated US data and tested on simulated US data and the worst performing network was the network trained on simulated US data and tested on real US data. Compared to the average dice coefficient (0.707) of the network trained using only limited real US data, the

transfer learning network had a lower average dice (0.586) and so did the mixed network (0.6760). Looking at the previous qualitative images, it can be seen that the Dice coefficient metric is not an ideal way of describing how well the automatic segmentation matches the manual segmentations. Therefore, it is also important to consider the AED values to consider how well the networks performed.

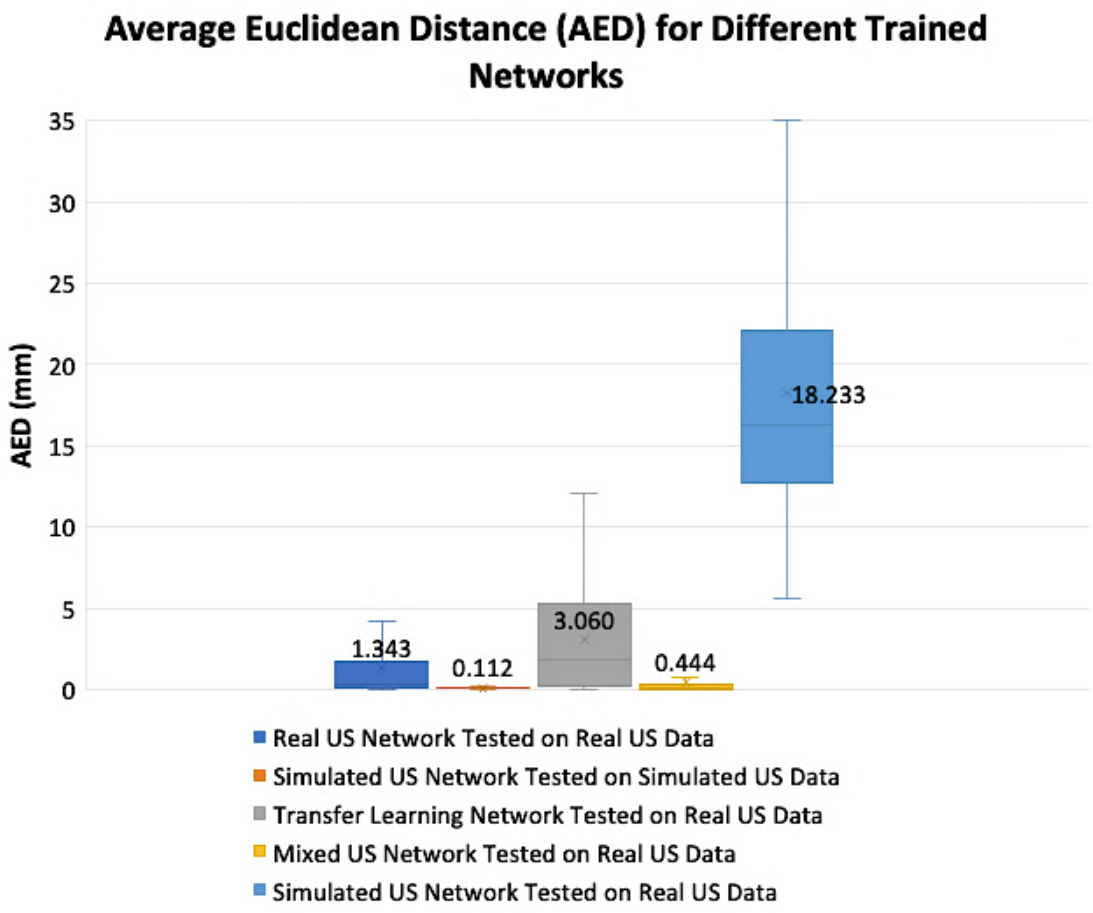


Fig. 3.7 This box and whisker plot shows the AED values for all of the different trained networks. The highest average AED value comes from the network trained on simulated US data and tested on real US data. The mixed network outperformed both the real US network and the transfer learning network.

The box and whisker plot in ***Fig. 3.7*** shows the AED values for the different networks. The simulated US network that was tested on simulated US data and the mixed network that was tested on real US data showed a very low variability in the AED values compared to the other trained networks that were trained. This demonstrates that the networks were trained well and had a good automatic segmentation results which closely matched the manual segmentations. The best and worst performing networks match the results provided by the dice coefficient values. The large-scale simulated US data network that was tested on simulated US data had the lowest average AED of 0.112 mm while the simulated US network tested on real US data had the highest average AED of 18.233 mm. In comparison to the average AED of the real US network, 1.343 mm, the transfer learning network had a higher AED of 3.060 mm while the mixed network had a very low average AED of 0.444 mm.

These results demonstrate that using simulated US data and applying transfer learning is not an efficient way of automatically segmenting the US bone surfaces. The dice coefficient and AED values of the transfer learning network fail in comparison to using only limited real US data. However, a mixed network developed using large-scale simulated data and limited real US data outperforms the network trained only on the real US data when comparing the AED value. Although the average dice coefficient value of the mixed network (0.676) was lower than the average dice coefficient value of the network trained on limited real US data (0.707), it was found that it is not a reliable metric for comparing bone surface segmentations the segmentations are not large regions that are ideally used for the dice coefficient calculations.

Therefore, the results show that large-scale simulated US data can be used in conjunction with limited real US data to train a neural network for accurate and robust segmentation of real i*n vivo* US data. Additionally, simulated US data alone cannot be used to segment real i*n vivo* US data due to the lack of artifacts and different regions that are found in the real US data.

3.8 Investigation of Validation Metrics

An evaluation of the validation metrics was done to demonstrate how well the dice coefficient and the AED show the difference between the manual and automatic segmentations of bone surface data. In order to perform this evaluation, simulated manual and automatic segmentations were generated and the values of the dice coefficient and AED were determined by varying either the distance or the length of the segmentations.

First, a vertical displacement analysis was performed. A manual segmentation was created which was a single line that was 5 pixels in thickness with a length of 151 pixels. Automatic segmentations were also created, and the segmentations were moved vertically downwards by 1 pixel and their corresponding dice and AED values were calculated. ***Fig. 3.8*** shows an example of the generated segmentations. The results that are shown in ***Fig. 3.9*** demonstrate that the dice coefficient decreases linearly as the automatic segmentation overlaps less rows of the manual segmentation. Although the segmentations may be close to each other in proximity, the dice coefficient will come out as 0 if there is no overlap between the regions of the manual and automatic segmentations. In contrast, the AED values show a different change where if the automatic segmentation only misses a couple rows of pixels, the AED value will stay towards the lower end. Due to the thin appearance

of bone surface segmentations from US data, automatic segmentations have a high likelihood of missing a couple rows of pixels when compared to the manual segmentations. However, if the segmentation is still overlaid on the bone surface then, the automatic segmentation would have been performed effectively but that cannot be seen from a low dice coefficient value.

a) b)

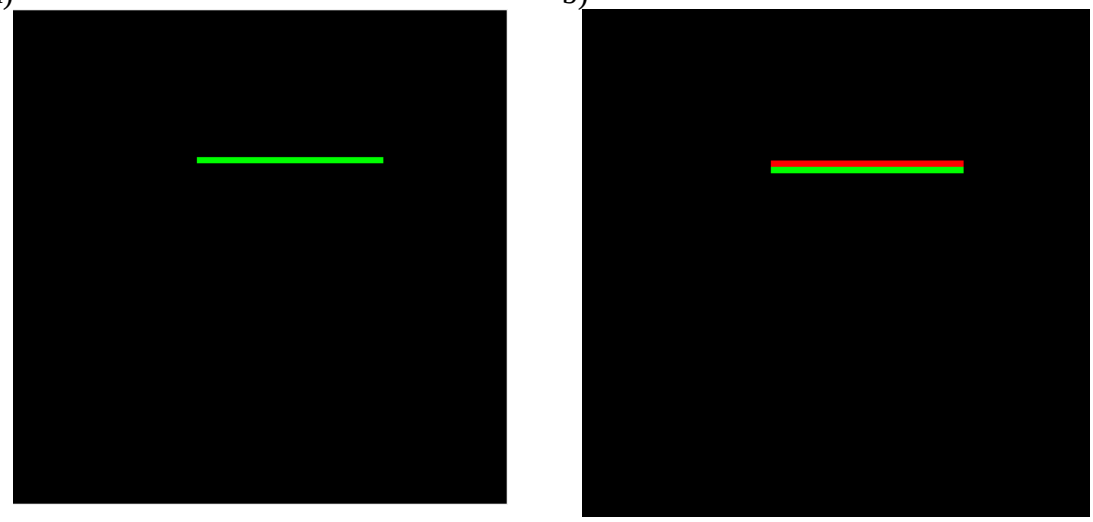


Fig. 3.8 Vertical displacement of a simulated manual segmentation (red) and automatic segmentation (green) was performed to evaluate the Dice and AED values. Images above show (a) complete overlap of the two segmentations and (b) complete mismatch of the two segmentations by translating the automatic segmentation down 5 rows of pixels.

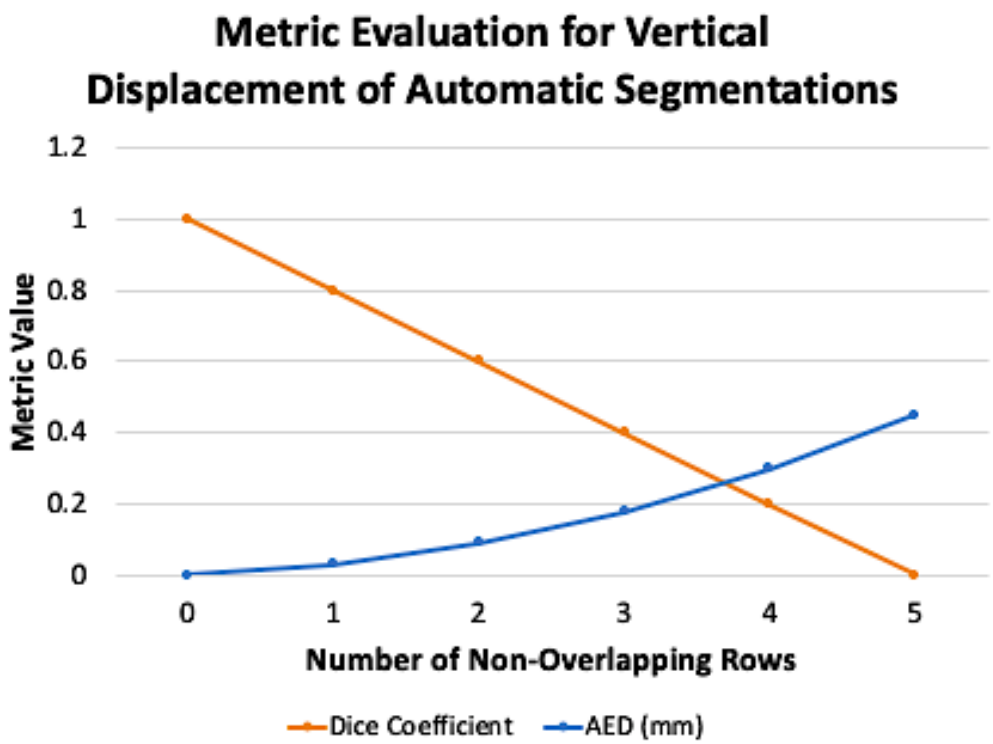


Fig. 3.9 A comparison of the Dice Coefficient and AED values for different vertical displacements of the automatic segmentations compared to the manual segmentation.

A horizontal displacement analysis was also performed to show how the dice coefficient and AED values change when part of the bone surface is not selected in the automatic segmentations. A manual segmentation was generated which was a line that was 5 pixels in thickness with a length of 151 pixels. The segmentations were then shifted down one row of pixels and the lengths of the automatic segmentations are decreased by 5 pixels at a time. The corresponding dice and AED values were then calculated. ***Fig. 3.10*** shows an example of the generated segmentations. The results that are shown in ***Fig. 3.11*** demonstrate if the automatic segmentation is one row lower and missing 25 pixels in length, the dice coefficient drops to a value of around 0.72. In comparison, the AED value for this error in the segmentation was found to be around 0.03 mm. The automatic

segmentation corresponding to this case can be seen in ***Fig. 3.10b***. The large drop in the dice coefficient value from 1 to 0.8 is due to the vertical displacement of 1 row of pixels. However overall, for horizontal displacement, the dice coefficient decreases slowly and the AED values increase slowly.

a) b)

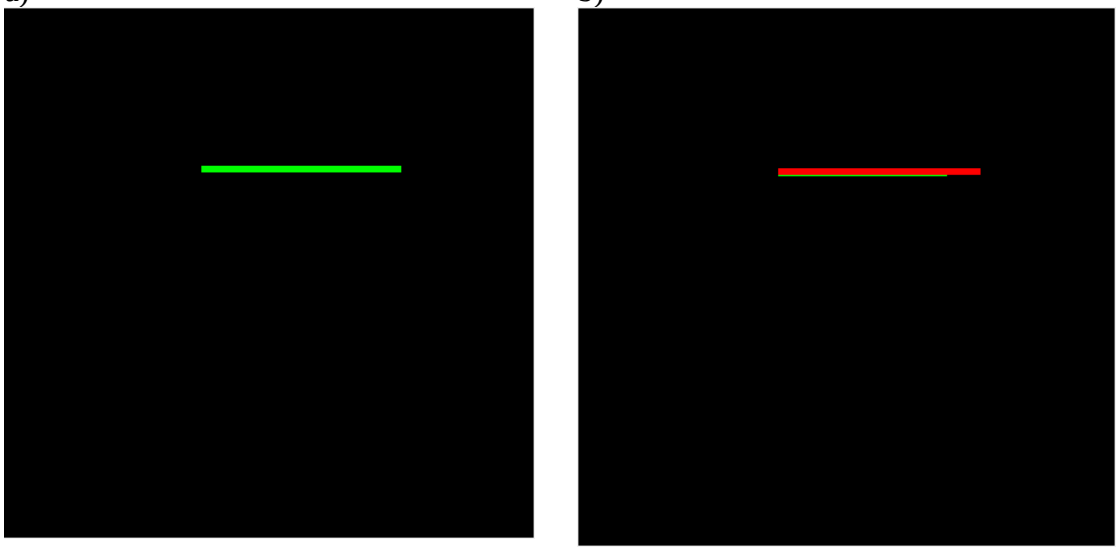


Fig. 3.10 Horizontal displacement of a simulated manual segmentation (red) and automatic segmentation (green) was performed to evaluate the Dice and AED values. Images above show (a) complete overlap of the two segmentations and (b) mismatch of the two segmentations by translating the automatic segmentation down 1 row and lessening the length by 25 pixels.

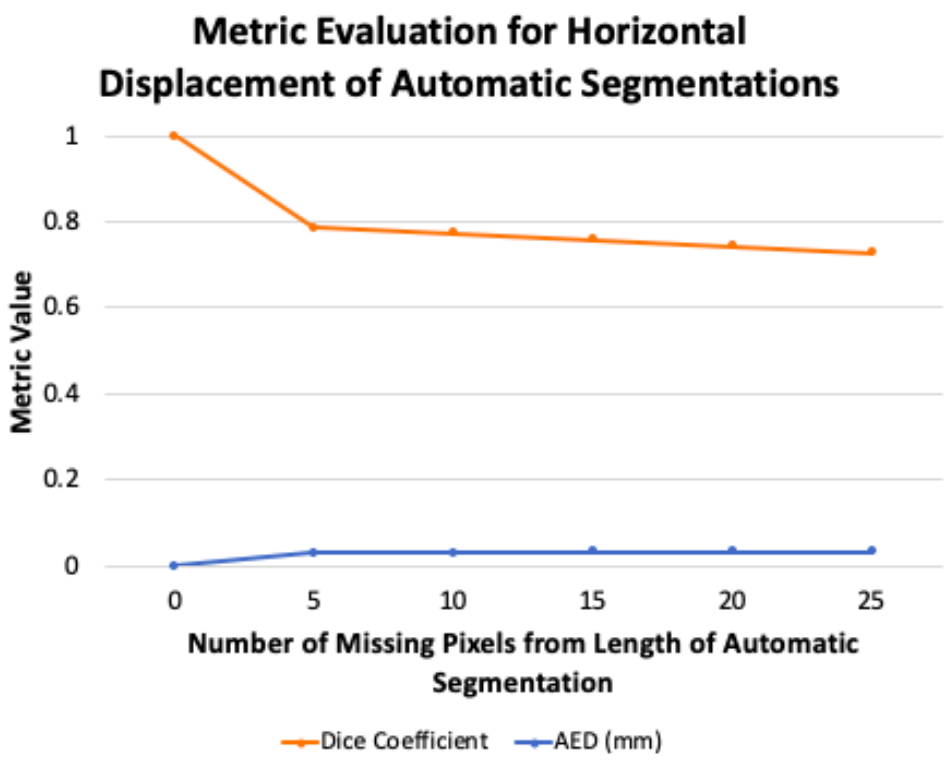


Fig. 3.11 A comparison of the Dice Coefficient and AED values for different horizontal displacements of the automatic segmentations compared to the manual segmentation.

The last evaluation of the validation metrics was performed to check how the metrics change if an additional surface is segmented in the automatic segmentation compared to the manual segmentation. To investigate this, a manual segmentation was generated that had only one surface outlined as the segmentation. Then, a corresponding automatic segmentation was generated that had the same surface selected as the manual segmentation, but it also had one additional surface segmented (***Fig. 3.12***). The dice coefficient of these two segmentations is 0.7139 while the AED was 16.4782 mm. This shows that the dice coefficient does not give an adequate value to show how the automatic segmentation fails if it has multiple regions segmented as a bone surface. This error in

automatic segmentation can occur when the algorithm incorrectly picks some of the high intensity values in the soft tissue interface as a bone surface.

a) b)

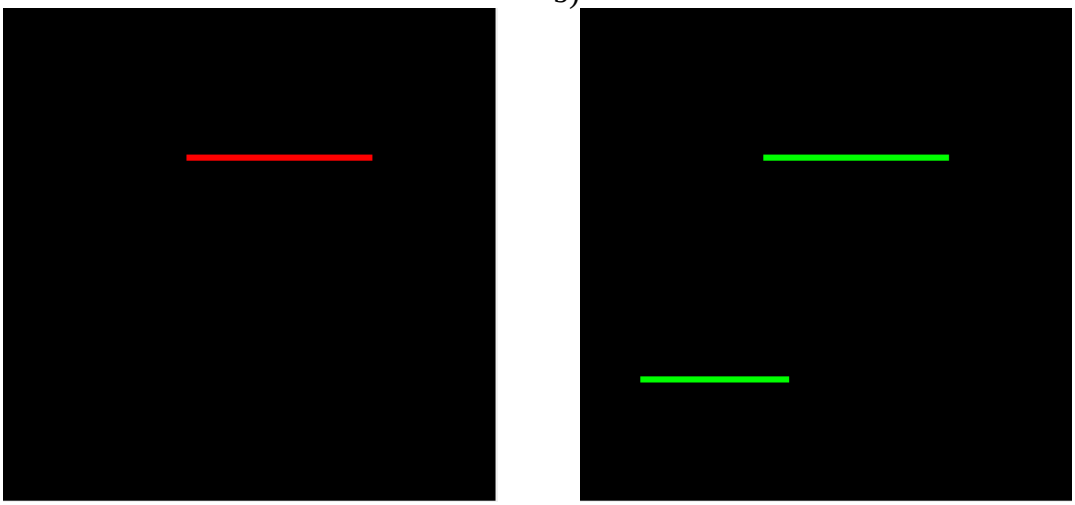


Fig. 3.12 a) shows the simulated manual segmentation with one bone surface segmented. However, image b) has an extra region segmented which can occur if the soft tissue is selected as a bone surface by the automatic segmentation. The dice value of these two segmentations is 0.7139 while the AED was 16.4782 mm.

CHAPTER 4

CONCLUSIONS AND FUTURE WORK

4.1 Significance of Research

In this thesis, a method of improving bone surface segmentation from limited real US data was investigated in order to determine a way to efficiently perform segmentations with a smaller real US dataset. The approach involved using large-scale simulated US data in conjunction with the limited real US data to train a network using different methods to determine if it would allow for segmentation of real US bone surface data.

Previous methods of segmenting real US bone surfaces have tried using traditional image processing and deep learning methods. However, the image processing methods have to overcome the problem of noise and artifacts in the US data. The use of traditional deep learning methods suffers from being dependent on a large amount of training dataset. In this thesis, it was proposed that using large-scale simulated US data along with limited real US data will help improve the accuracy of the deep learning-based automatic segmentation of bone surfaces from *in vivo* real US data.

Various training and testing strategies were investigated when training a well-known deep learning architecture developed specifically for segmenting medical data named U-net. First, U-net was trained using limited real US data and tested on real US data. Second, the U-net network was trained using large-scale simulated US data and was tested on both simulated and real US data. Third, a transfer learning approach was investigated to determine how it would affect the accuracy of segmenting real US data compared to only using real US data for training. In the transfer learning network, the U-net network was trained on large-scale simulated data and had the last layers retrained on

limited real US data. Fourth, the U-net network was trained from scratch with both large-scale simulated US data and real US data to determine how well it would segment real US bone surface data.

The results of the networks that used both large-scale simulated US data and real US data in their training were compared against the results of the network trained on limited real US data. This allowed for the determination of whether using large-scale simulated US data is a beneficial addition to only using limited real US data for performing automatic bone surface segmentations.

4.2 Contributions

As described at the end of Chapter 1.1, the work done in this thesis provides answers to the following two questions:

1) Can simulated US data be used to train a neural network for accurate and robust segmentation of real i*n vivo* bone US data?

   A neural network that was trained using large-scale simulated US data was able to perform accurate and robust segmentation of real *in vivo* bone US data when used in conjunction with limited real US data. A mixed neural network that was developed using both simulated and real US data outperformed the network that was trained using only limited real US data. However, the use of only simulated US data in a transfer learning network was not effective as the results of the dice coefficient and the AED showed that it did not have better values than the network that was trained on limited real US data.

2) Can simulated and real *in vivo* bone US data be used together to train a neural network for accurate and robust segmentation of real *in vivo* bone US data?

   Using simulated and real *in vivo* bone US data to train a neural network was shown to be effective as the results of the mixed network showed a smaller AED compared to the results of the network trained only on the limited real US data.

4.3 Limitations of this Thesis

This thesis explores the use of Simulated US data to help optimize a neural network to segment US bone surfaces. The Simulated US data that was obtained from 3D Slicer showed a bone surface image but with no artifacts present like in a real US scan which can also show soft tissue and other noise. The limitation with this data is that due to the lack of artifacts, the Simulated US images look similar even if the transducer was moved along the bone surface. Even when the transducer was rotated in 3D Slicer, only a small number of pixels changed because it was a slow rotational movement which led to the same bone surface with just a few different pixels. However, when the transducer was translated, the image moved accordingly but the shape of bone surfaces looked similar.

A problem with the real US data occurs when the manual segmentations are used for comparison. Manual segmentations, which are performed by expert annotators, can have both an inter-user and intra-user variability. Inter-user variability occurs when two different annotators segment the same bone surface US image but have slightly different segmentations. Intra-user variability occurs when the same annotator segments the same bone surface US image, but the segmentations have a variation. Therefore, a limitation of

this data is that those small errors, or variations, in the segmentations were not accounted for in the training and testing process. Although the variations might be small, some better method of validation which can incorporate those types of errors might help make the data more comprehensive.

4.4 Future Work

In order to effectively use the simulated US data that was used in the training of the neural networks, the simulated US data needs to be varied further. The data generated using 3D Slicer and PLUS used only one model of the human femur and radius. However, if different models can be found and used for the data (specifically models that come from different individuals) then the neural network training might give better results. Future work can include these different bone surfaces so that more scans be generated. A multi-institutional collaboration has been opened to gather more US scans and their segmentations so that they can be used in the future for developing better metrics for comparison and can improve the reliability of the manual segmentations [1].

Another aspect that can be improved is the method of validation that was used to compare the automatic segmentations to the manual segmentations. In this thesis, the Sorensen Dice Coefficient and Average Euclidean Distance were evaluated in order to determine how well the automatic segmentation matched the manual segmentation. The Sorensen Dice Coefficient is best used for larger regions and not optimal for a thin bone surface segmentation. Therefore, evaluation metrics that are based on the AED and how well the segmentations overlap would be better suited to give a good quantitative evaluation of the neural network.

Additionally, newer methods of segmenting bone surfaces have recently been explored. Neural networks can work in conjunction with image processing methods to further optimize the results obtained from the segmentations. Alsinan et al. have explored using B-mode US images and their corresponding local phase filtered images in order to incorporate them into a fusion network [2]. Another method that was utilized by Wang et al. used a network to enhance the bone surface present in the B-mode US scan and then used a modified version of the U-net in order to classify the bone surface and effectively segment the enhanced US bone surface image [3]. Newer methods have looked into using US elastography approaches to use US strain imaging and envelope power detection at each radiofrequency sample in order to localize the bone surfaces [4].

4.5 References

1. Pandey, P., Patel, H., Guy, P., Hacihalilogu, I., & Hodgson, A. J. (2019). *Preliminary Planning for a Multi-institutional Database for Ultrasound Bone Segmentation*. https://doi.org/10.29007/m1ll

2. Alsinan, A. Z., Patel, V. M., & Hacihaliloglu, I. (2019). Automatic segmentation of bone surfaces from ultrasound using a filter-layer-guided CNN. *International Journal of Computer Assisted Radiology and Surgery*, *14*(5), 775–783. https://doi.org/10.1007/s11548-019-01934-0

3. Wang, P., Patel, V. M., & Hacihaliloglu, I. (2018). Simultaneous Segmentation and Classification of Bone Surfaces from Ultrasound Using a Multi-feature Guided CNN. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. https://doi.org/10.1007/978-3-030-00937-3_16

4. Hussain, M. A., Hodgson, A., & Abugharbieh, R. (2014). Robust bone detection in ultrasound using combined strain imaging and envelope signal power detection. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. https://doi.org/10.1007/978-3-319-10404-1_45