Reliable and Robust Hip Dysplasia Measurement with Three-Dimensional Ultrasound and Convolutional Neural Networks

by

Houssam El-Hariri

B.A.Sc., Simon Fraser University, 2015

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

Master of Applied Science

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL
STUDIES
(Biomedical Engineering)

The University of British Columbia
(Vancouver)

March 2020

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the thesis entitled:

Reliable and Robust Hip Dysplasia Measurement with Three-Dimensional Ultrasound and Convolutional Neural Networks

submitted by **Houssam El-Hariri** in partial fulfillment of the requirements for the degree of **Master of Applied Science** in **Biomedical Engineering**.

Examining Committee:

Rafeef Garbi, Electrical and Computer Engineering
Co-supervisor

Antony J. Hodgson, Mechanical Engineering
Co-supervisor

Kishore Mulpuri, Orthopaedics
Supervisory Committee Member

Peter Cripton, Mechanical Engineering
Supervisory Committee Member

Abstract

Developmental Dysplasia of the Hip is one of the most common congenital disorders. Misdiagnosis leads to financial consequences and reduced quality of life. The current standard diagnostic technique involves imaging the hip with ultrasound and extracting metrics such as the α angle. This has been shown to be unreliable due to human error in probe positioning, leading to misdiagnosis. 3D ultrasound, being more robust to errors in probe positioning, has been introduced as a more reliable alternative. In this thesis, we aim to further improve the image processing techniques of the 3D ultrasound-based system, addressing three components: segmentation, metrics extraction, and adequacy classification.

Segmentation in 3D is prohibitively slow when performed manually and introduces human error. Previous work introduced automatic segmentation techniques, but our observations indicate lack of accuracy and robustness with these techniques. We propose to use deep Convolutional Neural Network (CNN)s for improving the segmentation accuracy and consequently the reproducibility and robustness of dysplasia measurement. We show that 3D-U-Net achieves higher agreement with human labels compared to the state-of-the-art. For pelvis bone surface segmentation, we report mean DSC of 85% with 3D-U-Net vs. 26% with CSPS. For femoral head segmentation, we report mean CED Error of 1.42mm with 3D-U-Net vs. 3.90mm with the Random Forest Classifier.

We implement methods for extracting α_{3D}, FHC_{3D}, and OCR dysplasia metrics using the improved segmentation. On a clinical set of 42 hips, we report inter-exam, intra-sonographer intraclass correlation coefficients of 87%, 84%, and 74% for these three metrics, respectively, beating the state-of-the-art. Qualitative observations show improved robustness and reduced failure rates.

Previous work had explored automatic adequacy classification of hip 3D ultrasound, to provide clinicians with rapid point-of-care feedback on the quality of the scan. We revisit the originally proposed adequacy criteria, and show that these criteria can be improved. Further, we show that 3D CNNs can be used to automate this task. Our best model shows good agreement with human labels, achieving an AROC of 84%.

Ultimately, we aim to incorporate these models into a fully automatic, accurate, reliable, and robust system for hip dysplasia diagnosis.

Lay Summary

The human hip is roughly a ball-and-socket joint. Some babies are born with hip dysplasia, which means the socket is less rounded than normal and the ball is less stable, which can lead to trouble walking and other problems. Ultrasound is used to scan the newborn baby for hip dysplasia, but doctors can still make mistakes using ultrasound. We would like to avoid mistakes, as mistakes lead to wasteful and potentially risky treatment options. In this thesis, we use 3-dimensional ultrasound and modern artificial intelligence techniques to assist clinicians in better diagnosing for hip dysplasia and making fewer mistakes.

Preface

The work in this thesis is part of the Developmental Dysplasia of the Hip (DDH) project, which was started by my three supervisors, Drs. Garbi, Hodgson, and Mulpuri, and builds on the work previously conducted by alumni of the lab including N. Quader and O. Paserin. The research done in this thesis was conducted in the Biomedical Signal and Image Processing Lab at UBC, the Surgical Technologies Lab at UBC, and the Orthopedic Research Lab at British Columbia Children's Hospital (BCCH). This work was approved by the UBC Clinical Research Ethics Board (CREB), certificate numbers: H14-01448, H18-00131, and H18-02024. Chapter 3 is based on the article listed below. The rest of the chapters are based on work that is not yet published.

Chapter 3 is based on the following paper, of which I was the first author, and which was reviewed and approved by the other authors:
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. In: Shen D. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. MICCAI 2019. Lecture Notes in Computer Science, vol 11765. Springer, Cham

Clinical data was collected at BCCH by our lab with the help of clinical technicians, nurses, and orthopedic surgeons. I lead data collection in 2019, and prior to 2019 data collection was lead by N. Quader and O. Paserin. I organized and labelled all the data. The studies conducted in this thesis were mainly designed by myself, with technical guidance from Drs. Hodgson and Garbi, and clinical guidance from Dr. Mulpuri. All technical work including coding, training and test-

ing of neural networks, clinical evaluations, and statistical analyses was done by myself. Drs. Hodgson and Garbi also helped reviewing this thesis.

Table of Contents

List of Tables

List of Figures

Glossary

2D 2-dimensional

3D 3-dimensional

ACA Acetabular Contact Angle

AI Artificial Intelligence

ANOVA Analysis of Variance

AROC Area under the Receiver Operating Characteristic curve

BCCH British Columbia Children's Hospital

BCE Binary Cross Entropy

BNN Bayesian Neural Network

CAE Centre Absolute Error

CAI Coverage Agreement Index

CAOS Computer-Assisted Orthopaedic Surgery

CDI Coverage Distance Index

CED Centre Euclidean Distance

CI Confidence Interval

CIHR Canadian Institutes of Health Research

CNN Convolutional Neural Network

CREB Clinical Research Ethics Board

CSPS Confidence-Weighted Structured Phase Symmetry

DDH Developmental Dysplasia of the Hip

DSC Dice-Sorensen Coefficient

FCN Fully Convolutional Network

FHC Femoral Head Coverage

FN False Negative

FOV Field-of-View

FP False Positive

GPU Graphics Processing Unit

HD Hausdorff Distance

HOG Histogram Of Gradients

ICC Intraclass Correlation Coefficient

ICICS Institute for Computing, Information and Cognitive Systems

IOU Intersection-Over-Union

LBP Local Binary Patterns

MC Multi-Channel

MED Mean Euclidean Distance

MRI Magnetic Resonance Imaging

MSE Mean Squared Error

NSERC Natural Sciences and Engineering Research Council of Canada

OCR Osculating Circle Radius

RAE Radius Absolute Error

REB Research Ethics Board

RFC Random Forest Classifier

RMS Root Mean Square

RNN Recurrent Neural Network

ROI Region-of-Interest

RQ Research Question

SD Standard Deviation

SOTA State-Of-The-Art

SP Shadow Peak

SPS Structured Phase Symmetry

TP True Positive

UBC University of British Columbia

US Ultrasound

VRMSE Vertical Root Mean Square Error

Acknowledgments

This thesis was funded by Natural Sciences and Engineering Research Council of Canada (NSERC), Canadian Institutes of Health Research (CIHR), and Institute for Computing, Information and Cognitive Systems (ICICS). The Titan V GPU was provided by NVIDIA. Compute Canada provided additional computing power and storage services.

I would like to thank my supervisors, Drs. R. Garbi, A.J. Hodgson, and K. Mulpuri for their support and guidance during the entire course of my studies at UBC. I would also like to thank my labmates for their support. I would like to thank the Orthopedic research team for their clinical support and for allowing us to share their space. I would like to thank the radiology technicians P. Thiessen, C. Beaton for their patience and help with data collection. I would also like to thank Dr. D. Rosenbaum (Radiology) for his support in helping me understand and interpret the ultrasound images. I would also like to thank Dr. A. Cooper for allowing us to scan his patients. I would also like to thank the nurses and the rest of the clinical team for their support on data collection days. Last, but not least, I would like to thank my family for financial and emotional support, and my friends for making this a fun experience.

Chapter 1

Introduction

Developmental Dysplasia of the Hip (DDH) is one of the most common congenital disorders affecting newborns [7, 16, 31, 42, 70, 71]. Accurate and early clinical diagnosis is key for effective treatment of DDH [78]. Current clinical practice for diagnosis usually involves Ultrasound (US) imaging of the newborn hip and manual delineation of anatomical landmarks. Despite its widely-accepted clinical use, there has been significant towards the low reliability of this 2D-US-based technique [32, 48, 66], which has recently motivated research towards using 3-dimensional (3D) US and automatic image processing as a more reproducible alternative to the current clinical practices. This thesis focuses on improving the image processing techniques to improve the reproducibility and robustness of 3D-US for measuring DDH.

1.1 Developmental Dysplasia of the Hip

1.1.1 Epidemiology

DDH is one of the most common congenital defects seen in newborns, with prevalence up to 2.85% [31], and incidence up to 7.6% in some populations [42]. Left untreated, DDH can lead to serious consequences including limping, leg length discrepancy, pain, and disability. Of particular note, DDH in infancy is a major risk factor in the development of early-onset osteoarthritis [10, 24, 25, 50, 69].

1.1.2 Diagnosis: Standard Clinical Practice

Due to the cartilaginous composition of the newborn hip (younger than 6 months), which is not easily resolved with X-ray, US is currently the clinical standard imaging modality for diagnosing DDH in newborns. In British Columbia, US examination is usually performed only if the infant is suspected of having DDH due to risk factors including: having been born breech; having been born with C-section; being female; and having a family history of DDH. US is usually performed in addition to a clinical examination that involves examining for leg length discrepancy; looking for hip folds; and applying the Barlow and Ortolani Tests to feel for dislocation clicks. X-ray is used for infants over 6 months of age, after the bone has begun to ossify.

Graf Ultrasound Technique

US-based testing was first popularized and standardized by Graf [12], who described the technique in detail. According to Graf, the newborn is placed in the lateral position, the hip is flexed, and the probe is positioned coronally at the hip joint. In this position, the sonographer would then look for the standard plane by navigating towards well-defined anatomical landmarks as shown in figure 1.1. When the sonographer identifies the standard plane, a 2D image is acquired. The sonographer then manually delineates salient anatomical landmarks and extracts two DDH measures: the α and β angles. The α angle is the angle between the ilium and acetabulum lines-of-best-fit, and measures the shallowness of the hip socket as shown in figure 1.2. A lower α angle indicates increased DDH severity. Similarly, the β angle is the angle between the ilium and labrum lines-of-best-fit. An increased β angle indicates increased DDH severity. Other measures have since been proposed including Femoral Head Coverage (FHC) originally proposed by Morin [47], which measures the percentage of the femoral head covered by the bony acetabulum as shown in figure 1.3. A decreased FHC indicates increased DDH severity. The normal to dysplastic ranges for the aforementioned DDH metrics are summarized in table 1.1.

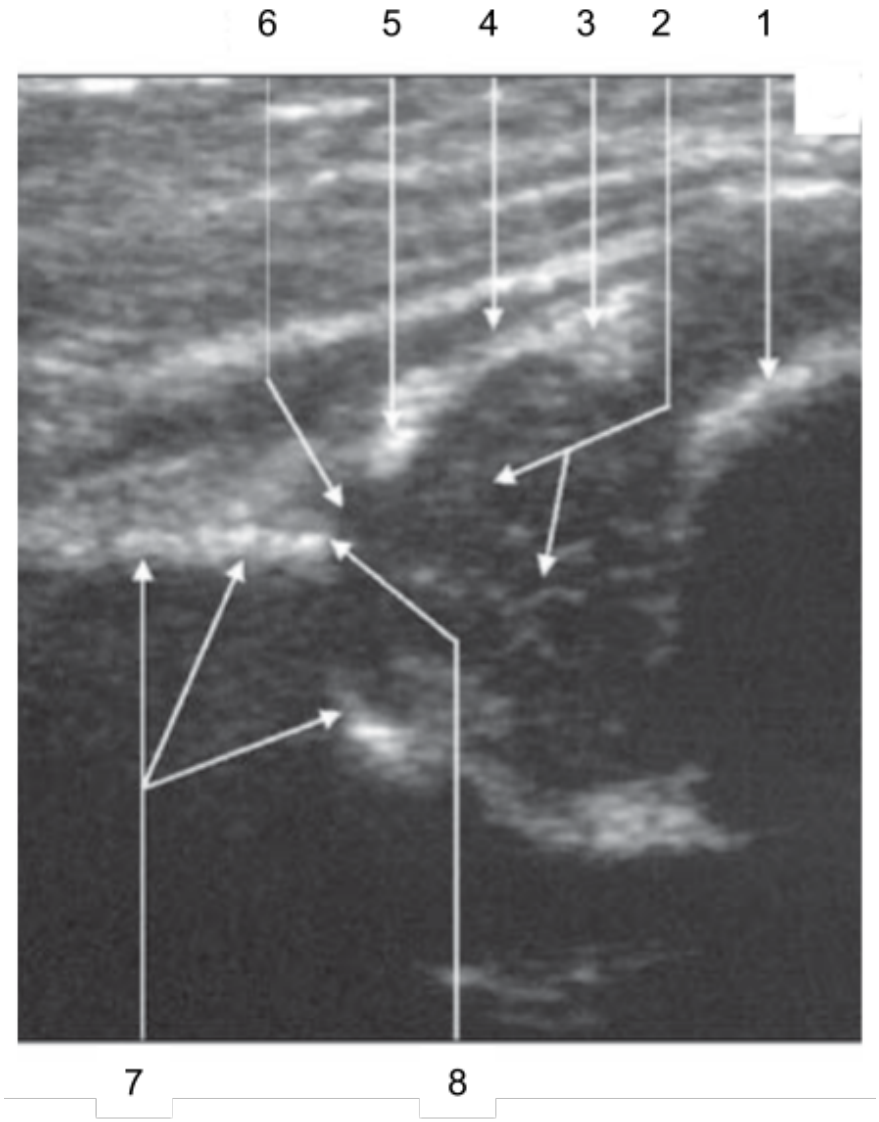


Figure 1.1: Ultrasound anatomical landmarks in the standard plane described by Graf [13]: 1) chondro-osseous junction; 2) femoral head; 3) synovial fold; 4) joint capsule; 5) acetabular labrum; 6) hyaline cartilage; 7) bony part of the acetabular roof; 8) bony rim: turning point from concavity to convexity.

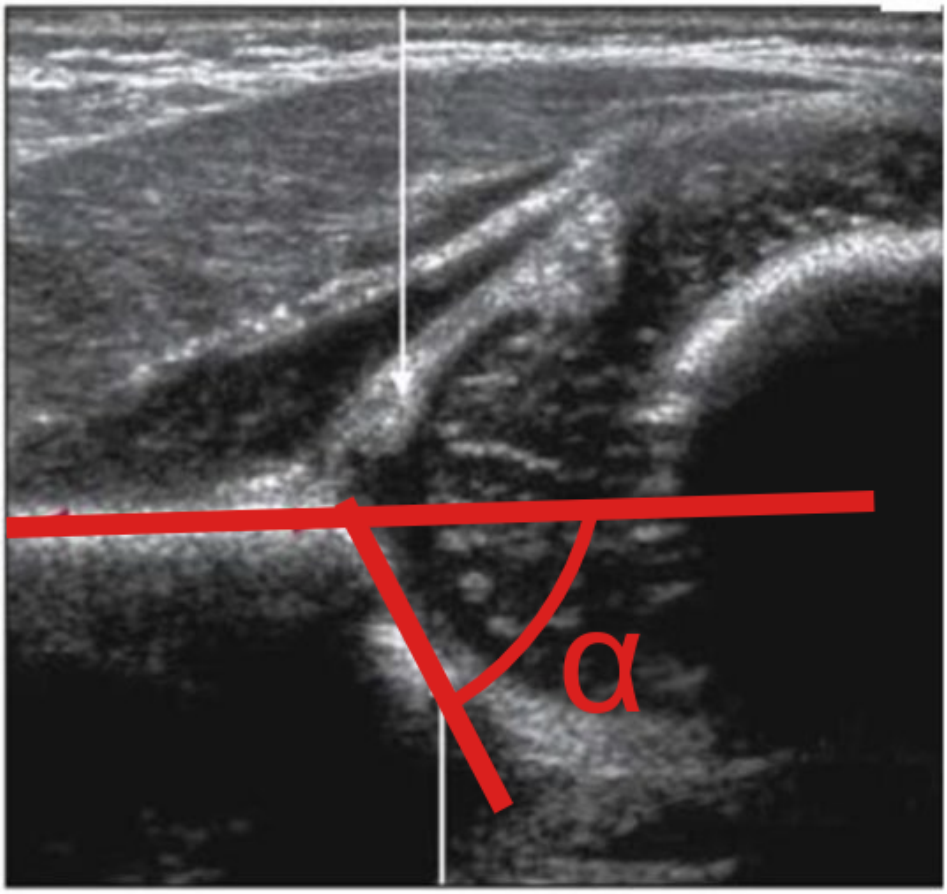


Figure 1.2: Illustration of the α angle proposed by Graf.

Table 1.1: Ranges of key DDH metrics

	Criterion, normal hip	Criterion, dysplastic hip	Range
α	$>60°$	$<43°$	17°
FHC	>55 %	<40 %	15 %

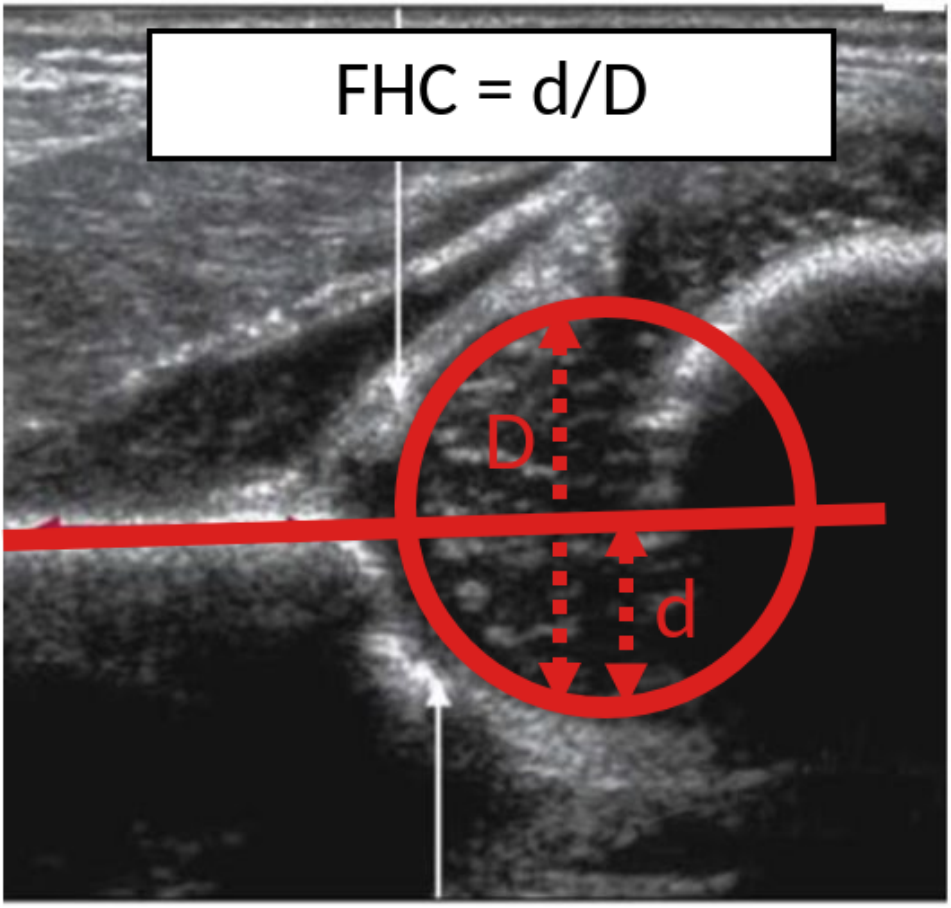


Figure 1.3: Illustration of Femoral Head Coverage proposed by Morin.

1.1.3 Treatment and Consequences of Misdiagnosis

If diagnosed early with DDH, most patients can be treated with the Pavlik Harness. For more severe cases, surgical intervention may be required. Dysplastic cases that were missed may require costly surgical interventions later in life. Over-treatment and under-treatment may result due to misdiagnosis, both of which have serious clinical and economic impacts on patients, their families, and society. We summarize potential consequences of misdiagnosis in figure 1.4, and present the following over-simplified discussion of costs only as a rough guide for the reader. Importantly, we do not report proportions of False Negative (FN)s and False Posi-

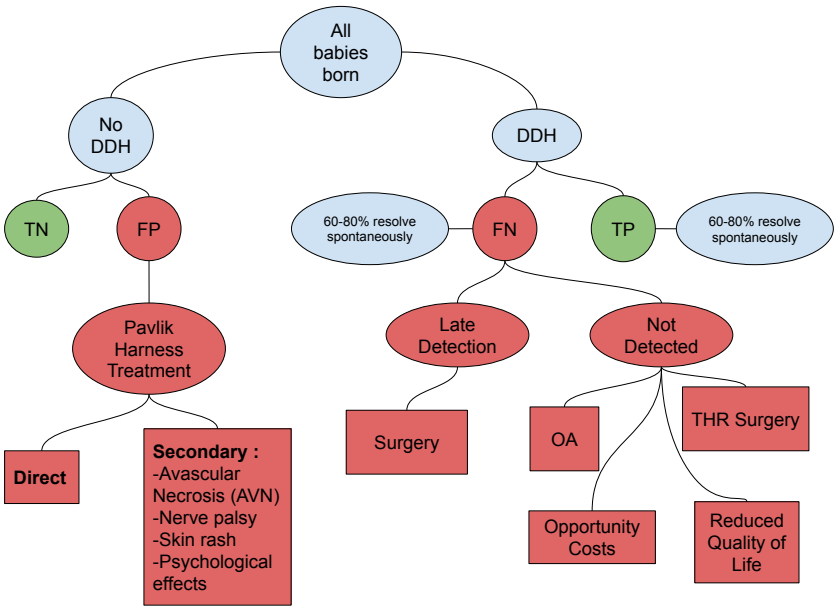


Figure 1.4: Simplified map showing socioeconomic consequences of DDH misdiagnosis in newborns. (TP: True Positive, TN: True Negative, FP: False Positive, FN: False Negative, OA: Osteoarthritis, THR: Total Hip Replacement)

tive (FP)s, as most sources in the literature attempt to quantify these against unreliable ground truth measurements such as 2D-US, which we consider an unreliable diagnostic measure as later explained. The analysis is further complicated by the fact that 60%-80% of cases born with DDH will spontaneously resolve within 2-8 weeks after birth [2]. Therefore, we leave a more in-depth analysis for future work.

False Negatives and Consequences of Under-Treatment

Newborns with DDH who are missed by standard diagnosis soon after birth may be detected in follow-up visits. As the Pavlik Harness loses efficacy 4 months after birth [51], more costly surgical treatments may be required. Consequently, the effect of late detection has been estimated to increase the cost of treatment by 7 times

due to late detection [78]. If completely undetected and untreated, DDH can lead to more serious consequences later in life, including early development of osteoarthritis, reduced quality of life, and opportunity costs. While it is difficult to quantify opportunity costs, there is some evidence towards the cost of osteoarthritis as a result of DDH. In a 2013 meta-analysis, Hoaglund reported an estimate that 10% of all osteoarthritis patients also had DDH [27]. However, it is worth noting that this is a conservative estimate and, in contrast, Nakamura [50] estimated that 88% of 2000 consecutive osteoarthritis patients in Japan had DDH. Taking Hoaglund's conservative estimate of 10%, Price [60] estimated that DDH might be responsible for about 25,000 hip replacements per year in the United States. At approximately $50,000 per procedure [68], the direct financial impact of these hip replacements is on the order of $1.25 billion per year in the United States alone.

False Positives and Consequences of Over-Treatment

Cases that are actually DDH negative but falsely identified as having DDH at birth also incur costs, and suffer serious consequences due to secondary complications. Direct costs include the cost of treatment with Pavlik Harness, reported to be on the order of £600 in the United Kingdom [78]. Given the relatively low cost and risk associated with use of the Pavlik Harness, erring towards over-treatment may seem a tempting option, but there are secondary factors against over-treatment to be considered. Avascular Necrosis, the death of bone tissues due to a lack of blood supply, is the worst and most frequent complication associated with the Pavlik Harness, reported in 1.35% to 10.9% of all infants undergoing treatment [40]. Even at such low incidence rates, this is a serious consequence that potentially requires very costly surgical treatment, and will cause much unnecessary suffering to the patient. Other secondary complications of Pavlik Harness include nerve palsy [49], skin rashes, and unnecessary psychological hardship on the parents [45].

1.1.4 Problem: Low Reliability of 2D Ultrasound for DDH Diagnosis

Although 2D-US is currently considered the clinical gold standard, there has been much evidence towards its lack of reliability. In a 2018 meta-analysis [61, 66], Quader reported extremely low inter-exam, inter-observer Intraclass Correlation

Coefficient (ICC)s for α angle of 23%, for β angle of 19%, and for FHC of near 0%.

Sources of Variability

A major source of error in 2D-US-based diagnosis is ambiguous probe positioning. In 2014, Jaremko showed by simulating US probe movement in three degrees-of-freedom that the α angle can vary by as much as 19° over the range of acceptable images (*acceptable* meaning it meets the definition of the standard plane) [32]. This is a very large range considering the 17° α angle range from extremely dysplastic at 43° to normal at 60° (see Figure 1.1). Consequently, a normal hip may be incorrectly diagnosed as dysplastic and vice versa. So, it appears that the operator's ability to identify the correct plane consistently is a major source of error and perhaps partially explains the extremely low ICCs.

Further exacerbating the problem is lack of standardization in training. Current clinical practice for diagnosing DDH with US involves identification of the standard plane as originally described by Graf [12, 13]. This is a difficult task and even experts with years of training can still make mistakes. For example, a 2013 paper by Graf on quality management of US hip sonography in Germany [14], reported that in 1.6%-43.7% of cases across 8 states the sonographers' licences were withdrawn by a quality control commission because of poor quality diagnosis. Further, Graf reported that in a refresher course, 250 orthopedic surgeons, pediatricians, and radiologists were required to classify 4 neonatal hip sonograms. Only 28% of the clinicians passed this test, and most mistakes were due tilting effects resulting in incorrect anatomical identification.

Another source of variability may be attributed to human variability in drawing the lines and circles used to make the measurements. However, this seems to be a relatively smaller source of error as evidenced by the high *intra*-image, inter-observer ICCs reported by Quader [61, 66] of 65%-90% for α and 70%-93% for FHC.

Therefore, it appears that the main problem leading to variability is high variability in probe positioning.

1.1.5 Solution: 3D Ultrasound

To mitigate human variability in probe positioning, and improve diagnostic accuracy and reliability of DDH, 3D-US is the most promising solution that has been proposed in the last few years.

Quader proposed in 2016 the α_{3D} angle, a novel 3D-US-based metric [63] analogous to the α angle originally proposed by Graf [12]. They hypothesized that 3D-US is more robust to ambiguous probe positioning, and would improve test-retest reproducibility over 2D-US [61]. Quader's implementation was based on Confidence-Weighted Structured Phase Symmetry (CSPS) for segmenting the pelvis bone surface and was fully automatic, not requiring operator input to segment the bone surface. They reported a 75% reduction in test-retest Standard Deviation (SD) with α_{3D} compared to α [61].

Concurrently in 2016, Mabee and Hareendranathan proposed another 3D-US-based metric, the Acetabular Contact Angle (ACA) [21, 44]. Their implementation of ACA was semi-automatic and required some human input. The reported inter-exam, intra-rater variability for ACA was 41%, reported in Quader's meta-analysis [61, 66] as test-retest SD normalized over the range of angles from normal to dysplastic.

Quader later proposed in 2017 FHC_{3D} [64], a novel 3D-US-based metric analogous to FHC originially proposed by Morin [47]. FHC_{3D} was defined as the the ratio of the femoral head volume medial to the Ilium plane-of-best-fit vs. the total volume of the femoral head. Quader's implementation was fully-automatic, and resulted in a 65% reduction in test-retest SD compared to the analogous FHC 2D-US-based measure [61].

Most recently, Zonoobi and Hareendranthan proposed α_{3D}-posterior, α_{3D}-anterior, and Osculating Circle Radius (OCR) 3D-US-metrics [82]. Zonoobi reported inter-exam ICCs of 68%, 62%, 50% for α_{3D}-posterior, α_{3D}-anterior, and OCR, respectively. Using the same techniques, Mostofi conducted a study in 2019 [48] to compare the benefit of 3D-US in the hands of novice (1.5 hrs. training) vs. expert (5 years training) users. For novice users, Mostofi reported inter-exam α angle ICC of 10% vs. α_{3D} angle ICC of 73%-83%, showing that with 3D-US novices can measure DDH almost as consistently as experts whereas this is not the case with 2D-US.

These works have shown substantial evidence in support of the hypothesis that 3D-US-based DDH diagnostic metrics are more reliable as compared to 2D-US-based metrics. To the best of our knowledge, we are not aware of any other works in the literature on automatic 3D-US for DDH. Further, we acknowledge that there have been many works on 2D-US for DDH, which we do not address in this thesis due to the strong aforementioned evidence against this imaging modality for hip dysplasia measurement.

1.2 Overall Objective

The ultimate goal of this project is to develop a system for safe, reliable, accurate, and robust measurement of DDH, and that is amenable for clinical translation. US, being a safe, non-ionizing modality; capable of imaging the cartilaginous hip joint anatomy; and being a relatively portable and more cost-effective technology, is the imaging modality of choice for our system. Specifically, we choose 3D-US due to being more reliable than 2D-US. Therefore, we aim to develop a system that can reliably and robustly extract key DDH metrics such as the α_{3D} angle from 3D-US images of the neonatal hip. A simplified workflow of our envisioned system is depicted in figure 1.6. In the next section, we present preliminary technical work that has been done by our lab and others towards such a system, and we propose areas of improvement that will drive this thesis. Specifically, we address key steps in the pipeline including:

- Automatic, accurate, and fast segmentation of salient anatomy including the pelvis bone surface and femoral head
- Automatic, reliable, and robust key DDH metrics extraction from the segmented volume
- Accurate and fast adequacy classification of images to provide point-of-care feedback to the sonographer about the quality of the acquired image

1.3 Related Work

1.3.1 Segmentation

Evidently, much progress has been made with 3D-US for DDH diagnosis and measurement, but there are still some aspects that can be improved which would help facilitate clinical adoption of the proposed system. For example, Hareendranathan's graph-based segmentation method is *semi*-automatic [21, 23], which unnecessarily requires additional clinical time for seed-point entry (30-75 seconds [82]); introduces another source of variability; and could be a barrier to clinical adoption especially in settings where trained US technicians are scarce. Quader's methods on the other hand, are fully automatic. Hand-crafted features are used including bone shadowing and phase symmetry to segment the pelvis bone surface [61, 62]. To segment the femoral head, they use multiple Random Forest Classifier (RFC)s, which take as inputs many features including Histogram Of Gradients (HOG) and Local Binary Patterns (LBP) [61, 64]. However, our recent observations using Quader's algorithm show dubious behaviour, including most commonly (see figure 1.5):

- Over-segmentation of soft-tissue
- Under-segmentation of the pelvis bone surface, in some cases missing most of the pelvis bone surface and capturing only a thin sliver
- Under-segmentation of the femoral head
- Incorrect plane-fitting due to noisy segmentation

Ultimately, leading to concerns about the validity of the α_{3D} and FHC_{3D} measures, despite being consistent between scans. For these reasons, part of this thesis focuses on fully automatic methods for segmentation and localization of the salient anatomy in neonatal hip 3D-US, in an effort to improve DDH measurement. The next sections present a discussion of recent related work in US bone segmentation.

Bone Segmentation in Ultrasound

Automatic segmentation of bone surfaces in US is a well-studied area, with the majority of work having been focused on Computer-Assisted Orthopaedic Surgery (CAOS) applications in *adult* patients [17, 53]. Traditionally, a variety of techniques based on hand-crafted features have been explored including intensity-based analysis, morphological operations, connected-component analysis, phase analysis, and others [17, 53]. In recent years, we have seen major successes of deep-learning-based techniques for segmentation in medical imaging and other areas, and naturally we have observed a similar trend for bone segmentation in US in the last year.

Several works have proposed solutions for bone surface segmentation in US, tested on *adult* bone US. For example, Villa proposed using the popular FCN-8s architecture [43], concatenating the B-mode image with phase symmetry (PS) and bone shadowing features in the input channels, and compared the performance of this multichannel approach to the CSPS approach proposed by Quader [62]. Villa [74] reported a DSC of 57%±28% for the multichannel Fully Convolutional Network (FCN) compared to 41%±25% for CSPS, showing a significant improvement in accuracy. Wang [76] proposed using another popular architecture, U-Net [67], and similarly fusing with the input B-mode image features including Bone-Shadow Enhanced Image, Local Phase Tensor Image, and Local Phase Bone Image. Wang reported a DSC of 97% with this approach, and 93% with vanilla B-mode U-Net. In a more recent work by the same group, Alsinan [1] proposed a new architecture based on AdapNet [73], with late-stage Local Phase feature fusion, and reported DSC of 98%, and 91% with vanilla B-mode U-Net.

Bone Segmentation in Neonatal Hip Ultrasound

Several works have explored automatic segmentation specifically in *neonatal hip* US for DDH diagnosis, which presents unique challenges due to the partially cartilaginous composition of neonatal bone.

Hand-engineered phase features were proposed, of which a prominent example is the aforementioned CSPS technique proposed by Quader [61, 62]. CSPS combined Structured Phase Symmetry, an orientation-independent variant of Phase

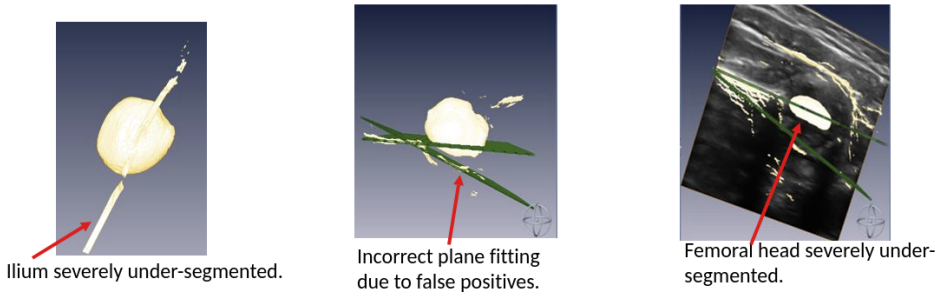


Figure 1.5: Visualizing performance of Quader's methods. Problems shown include over-segmentation, under-segmentation, and incorrect plane-fitting.

Symmetry [39] designed to segment non-planar bone structures, with bone shadowing features reducing soft tissue false positives. More recently, Pandey [54] proposed Shadow Peak (SP), a simplified method that uses only bone shadowing features to segment bone, and has shown certain improvements in accuracy and speed over CSPS in a limited study. Though promising, these methods still rely on highly engineered hand-crafted features hence challenges remain with regards to robustness and generalizability to new data, as we later show.

Data-driven methods have also been proposed for this task. Hareendranathan proposed using superpixel classification with a CNN and reported Hausdorff Distance (HD) error of 2.1$\pm$0.9mm between contours [22]. Zhang [81] proposed a neural network based on Mask R-CNN, and compared it to other the popular architectures including FCN-32s and U-Net but they reported very poor DSCs of 39% for their network, 5% with U-Net, and 22% with FCN-32s (we note that their results contradict our own tests and findings, as will be presented later). Golan [11] applied U-Net with an extra adversarial component for automatically segmenting the ilium and acetabulum bone surfaces for automatically extracting α angle from the 2D coronal standard plane. Golan did not directly report segmentation agreement with human labels, but reported a correlation coefficient of 0.76 with the clinical α angle.

1.3.2 Adequacy Classification

Addressing the *adequacy assessment* step (figure 1.6), Quader [61, 65] presented the first work on adequacy assessment of 2D-US for DDH. Quader's method relied on extracting certain features including HOG and LBP, and using a RFC to classify whether coronal slices are adequate for measurement. They reported an excellent AROC for this technique of 98.5%. Paserin conducted the first work on adequacy classification of neonatal hip 3D-US *volumes* [56–58]. The goal of this work was to implement a classifier that could provide rapid point-of-care feedback to the operator whether the acquired volume is adequate for measurement or must be re-acquired. The advantages here are improved workflow efficiency and speed, as the existing methods for automatically extracting DDH metrics from 3D-US were relatively slow, requiring on the order of 1 minute computation time, thus processing an inadequate image would be a waste of valuable clinical time. Further, such a classifier could reduce costs by helping inexperienced users in remote locations to scan patients locally. For example, currently patients in Canadian Territories are flown to British Columbia for DDH examination due to lack of expertise in these remote locations.

Paserin first proposed using a Convolutional Neural Network (CNN)-based adequacy classifier [57], that processes volumes frame-by-frame, and uses an aggregate score to classify the full volume. Specifically, the CNN was trained to look for coronal slices containing all anatomical landmarks including the ilium, acetabulum, femoral head, ischium, and labrum. Paserin reported a per-slice cross-validation classification accuracy of 90% and runtime of 2s per volume [56], much faster than Quader's method [65] which requires around 3 minutes per volume. Paserin later also proposed the addition of a Recurrent Neural Network (RNN) [58], which similarly processes coronal slices frame-by-frame, but can make use of information in adjacent slices. Paserin hypothesized that incorporating this information would improve the classification accuracy. With this model, Paserin reported an improved per-slice accuracy of 93% [56]. To the best of our knowledge, there were no other works on automatic adequacy classification of DDH 3D-US volumes outside of our group.

Despite showing excellent accuracy in terms of agreement with expert human

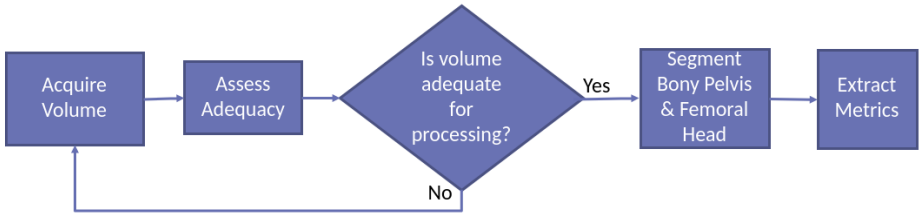


Figure 1.6: High-level conceptual design of our system.

labels on a limited dataset, as well as improvements in computation time, Paserin's work [56, 58] did not evaluate the *choice of adequacy criteria*. In this thesis, we revisit these criteria and present modified criteria and new methods for automatic adequacy classification.

1.4 Research Questions Addressed

Some obstacles remain for clinical translation of the 3D-US-based system for DDH measurement. Hareendranathan's methods [23, 82] are *semi-automatic*, requiring user input of key points to segment the pelvis bone surface and femoral head. Quader's methods [61], on the other hand, are fully automatic and use CSPS for segmenting the pelvis bone surface, and an RFC to segment the femoral head. However, our recent observations show that these methods which rely on highly-engineered, hand-crafted features often fail to correctly segment the salient anatomy for DDH measurement. The last few years have witnessed the rise of deep Convolutional Neural Network (CNN)s, which consistently showed overwhelming evidence for their ability to outperform classical machine learning methods that rely on hand-crafted features for image processing. Our first research questions are as follows:

- Research Question 1: Can CNNs be trained to segment the *pelvis bone surfaces*, including the ilium and acetabulum, in neonatal hip 3D-US? Would the predictions produced by such CNNs more closely resemble human labels, as compared to existing SOTA methods such as CSPS?

- Research Question 2: Can CNNs be trained to locate the *femoral head* in neonatal hip 3D-US? Would the predictions produced by such CNNs more closely resemble human labels, as compared to existing SOTA methods such as Quader's RFC?

Several automatic techniques have been proposed for extracting DDH metrics from segmented neonatal hip 3D-US volumes. However, these methods were highly tailored to the segmentation techniques previously proposed. For example, Quader developed highly engineered techniques to get α_{3D} [61, 63] from the relatively noisy CSPS segmentation of the bone surface. Perhaps these techniques could be further improved, given improved segmentations, and can produce more reproducible, robust, and plausible DDH diagnostics.

- Research Question 3: Can we develop automatic methods for extracting α_{3D} and FHC_{3D} metrics with our improved segmentations that are at least as reproducible as the previously proposed methods [61, 63, 64]? Can we show that our proposed methods are at least as robust and plausible as these previously proposed methods?

Paserin first attempted to standardize and automate neonatal hip 3D-US adequacy classification based on anatomical landmarks [56–58]. Paserin showed that CNNs are capable of rapidly and accurately predicting adequacy based on predefined criteria. However, Paserin did not explore the completeness of the adequacy criteria and how they relate to DDH measurement.

- Research Question 4: Are the current adequacy criteria proposed by Paserin [56] sufficient? Can we improve the criteria? Can we train new models for automating classification based on the newly defined criteria?

1.5 Contributions

Our contributions are as follows:

- As part of our clinical study and for training neural networks, we collected 3D-US data from 59 newborn participants at BCCH, further expanding this project's database.

- Training, comparing, and evaluating U-Net [67] and 3D-U-Net [6] for segmenting pelvis bone surface in 3D-US. Towards this end, I labelled ∼600 2D slices and ∼100 3D volumes. We show much improved segmentation accuracy with these methods compared to the SOTA.

- Training, comparing, and evaluating CNNs including 3D-ResNet-50 [19, 20] and 3D-U-Net [6] for segmenting and locating the femoral head in neonatal hip 3D-US. In the process, I labelled ∼100 volumes for training and evaluating the neural networks. We show much improved segmentation and localization accuracy with these compared to the SOTA.

- Proposing new algorithms for extracting α_{3D} [63], FHC_{3D} [64], and OCR [82]. We show improved reliability, robustness, and plausibility with the proposed algorithms compared to the SOTA.

- Revisiting the adequacy criteria previously proposed by Paserin [56–58], and evaluating the the choice of these criteria. We propose new criteria, and show that 3D-CNNs can be trained to automate adequacy classification based on the newly proposed criteria. We show that the new criteria are more selective and this selectivity improves test-retest reproducibility of DDH measurement.

1.6 Thesis Outline

In addition to this introductory chapter, this thesis includes six chapters, outlined as follows:

- Chapter 2: Presents an overview of the clinical study, details of the data collected, terminology, and conventions used in this thesis.

- Chapter 3: As per RQ-1, we train U-Net [67] for pelvis bone segmentation in neonatal hip 3D-US, and evaluate its performance against SOTA methods for pelvis bone segmentation including CSPS [54, 61].

- Chapter 4: We address RQs 1, 2, and 3.

  - Section 4.1: We leverage U-Net from Ch.3 to train 3D-U-Net [6] to further improve the segmentation of the pelvis bone surface.

  - Section 4.2: We apply 3D-CNNs to segment the femoral head, and compare our models' performance to Quader's SOTA RFC [64].
  - Section 4.3: We use our improved segmentation of the pelvis bone surface and femoral head to implement a new algorithm for extracting α_{3D}, FHC_{3D}, and OCR. We evaluate our metrics compared to Quader's [61] and Hareendranathan's [82], and in the process present new adequacy classification criteria.

- Chapter 5: This chapter focuses on revisiting adequacy classification criteria previously proposed by Paserin [56], and presenting new adequacy criteria. We assess the effect of using the old and new criteria on DDH measurement. Finally, we evaluate 3D-CNNs for automating the task of adequacy classification.

- Chapter 6: Final discussion of conclusions, limitations, and future work.

Chapter 2

Clinical Protocol and Description of Data

US data from real participants is a key component for the work presented in this thesis, so we dedicate this chapter to describing how the data was collected and used. As part of an ongoing effort since 2014, we collected data from a total of 118 participants at British Columbia Children's Hospital (BCCH) (Vancouver, British Columbia, Canada), with the collaboration of engineering students, professors, orthopedic surgeons, radiologists, US technicians, nurses, and research staff. Data was collected with the approval of the UBC Children's and Women's Research Ethics Board (REB), under the following application IDs:

- Phase I (2014-2018): H14-01448
- Phase II (2018-2019): H18-00131
- Phase III (2019-Present): H18-02024

Roughly the same protocol was used over the three phases for collecting 3D-US from newborn participants, with some small differences due to changing research questions and clinical workflow over the years, which will be briefly described in this chapter.

Figure 2.1: Ultrasonix 4DL14-5 3D ultrasound probe.

2.1 Inclusion and Exclusion Criteria

Participants were selected based on the following criteria.

Inclusion criteria:

- Suspected or diagnosed with DDH, and this is usually due to other risk factors including being born with C-section, born breech, being female, or having a family history of DDH.
- Referred for a regular ultrasound exam
- Ages 0 to 6 months of age (0-4 months in Phases I and II).

Exclusion criteria:

- Not suspected of having DDH
- Not referred for clinical US exam
- Has other congenital hip abnormalities
- Over the age limit

2.2 3D-US Data Collection Protocol

When a newborn is suspected of having DDH, they are normally referred for a clinical US exam. With the parents' consent, 3D-US scans for our study are collected after the technician scans the newborn with a 2D probe as part of their regular clinical practice. The Ultrasonix 4DL14-5/38 (BK Ultrasound, Richmond, BC, Canada, see figure 2.1), which uses a mechanically-swept 1D linear piezoelectric array to get 3D images, is used to acquired 3D images in this study. For each participant, the sonographer attempts to scan each hip twice, with the probe being removed and replaced between exams. Before starting the recording, the sonographer is asked to align the probe in the coronal plane, finding the optimal plane according to their training. With the probe being held as steady as possible in this position, multiple sweeps of the array are acquired within an exam. Using this protocol, multiple volumes are acquired per participant (varies depending on the protocol followed in each phase, and the participant's level of cooperation).
We note the following differences between the phases:

- Phase I:
  - Orthopedic surgeons did the scans
  - Multiple surgeons imaged each participant
  - Static assessment protocol was followed, so no force was applied to the participant's hip
  - Number of sweeps per exam varied up to 8
- Phase II:
  - Specialized US radiology technicians did the scans
  - Only one technician imaged each participant
  - Dynamic assessment protocol was followed [59], so the participant's hip was pushed posteriorly usually in the second and third sweeps in an exam
  - 4 sweeps were usually collected per exam
- Phase III (see Figure 2.2):

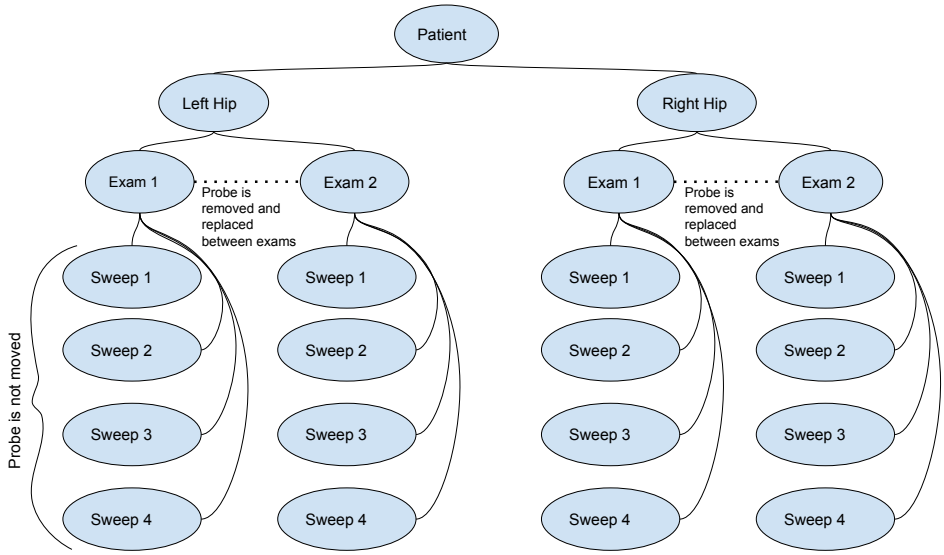


Figure 2.2: Data collected per patient in Phase III, showing number of exams per patient and number of sweeps per exam.

- Specialized US radiology technicians did the scans
- Only one technician imaged each participant
- Static assessment protocol was followed, so no force was applied to the participant's hip
- 4 sweeps were usually collected per exam

2.3 Data Used in Each Chapter

The data used in the following chapters is summarized in figure 2.3. The total number of participants is 118, and the total number of sweeps from all participants is 2,202. In Chapters 4 and 5, to ensure as many participants as possible are represented in the dataset given the labeling time constraints, we down-sample by randomly selecting only 4 sweeps per participant, so we end up with 472 *sweeps* which is reflected in the x-axis of figure 2.3. The numbers on the right hand side represent the number of *participants* per category. We briefly describe the data used in the subsequent chapters here:

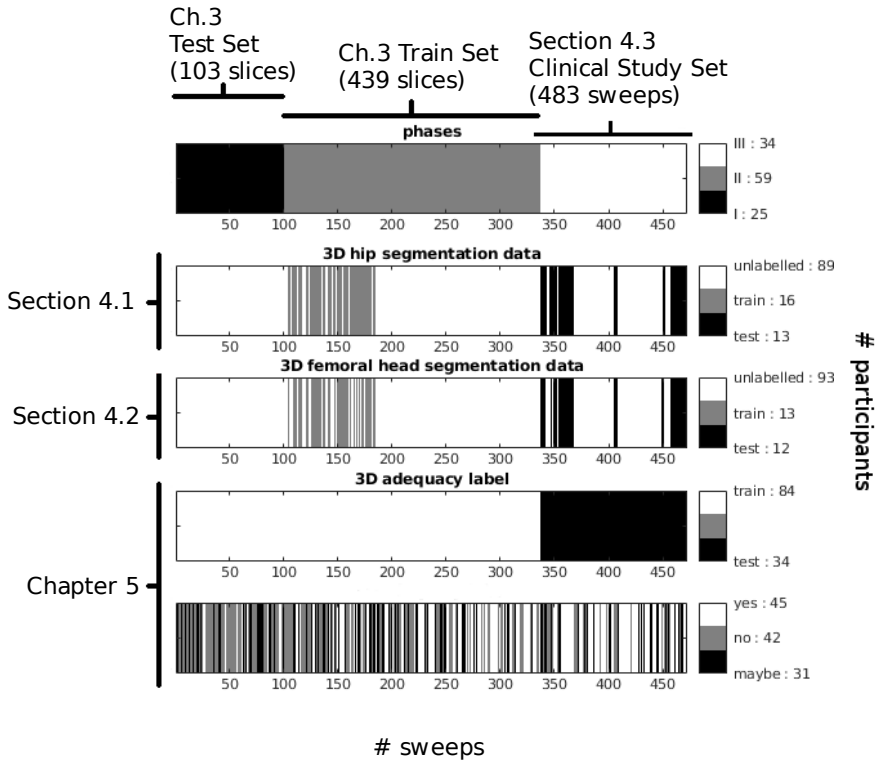


Figure 2.3: Summary of 3D ultrasound data we collected from BCCH, showing details of training and testing data used for each chapter in this thesis.

- Ch.3: The data in this chapter was used to train and test U-Net, which processes volumes slice-by-slice. Briefly, we randomly extract 439 coronal slices from the Phase II sweeps for training, and 103 coronal slices from the Phase I sweeps for testing. Not shown in figure 2.3, we additionally collected 72 2D coronal US images with the Clarius L7 wireless probe in Phase III, which we use to evaluate U-Net's performance on unseen data from a different probe.

- §4.1: 3D-U-Net is used, so we require 3D binary mask labels. We label 64 sweeps from Phases I and II for training, and 52 sweeps from Phase III for testing.

- §4.2: Again, we use CNNs that use 3D convolutions, so require 3D labels. We label 52 sweeps from Phases I and II for training, and 48 sweeps from Phase III for testing.

- §4.3: For the clinical study described in this section, we make use of all 483 sweeps in Phase III before being down-sampled, as in this case the labeling involved is much faster than the previous chapters. As we illustrate in figure 2.2, for each participant both hips are imaged. Each hip is scanned twice, with the probe being removed and replaced to assess test-retest reproducibility. Within each exam we usually record 4 anterior/posterior sweeps. This sums to 16 sweeps per participant, however this number could vary depending on the level of cooperation from the participant.

- Ch.5: Similar to §4.1 and §4.2, we again require labeled 3D data to train 3D-CNN adequacy classifiers. Since it requires much less time to get the yes/maybe/no labels for this task compared to the pixel-wise segmentation labels required in the previous chapters, we were able to label the full set of 472 sweeps. We assign all sweeps in Phases I and II to the training set and all in Phase III to the test set.

2.4 Coordinate System Description

We can describe a 3D-US volume in terms of anatomical terms, Cartesian coordinates, or matrix coordinates, so in this section we describe the relations between these coordinate systems to clarify terms used in the rest of the thesis. The conventions used are illustrated in figure 2.4. Anatomical terms are shown in white, Cartesian coordinates in purple, and matrix terms in black. To conform with clinical practice, in this thesis we always show hip US in the horizontal-cranial left position [14]. Matrix *rows* are aligned with the y-axis; matrix *columns* are aligned with the x-axis; *coronal slices* are aligned with the z-axis.

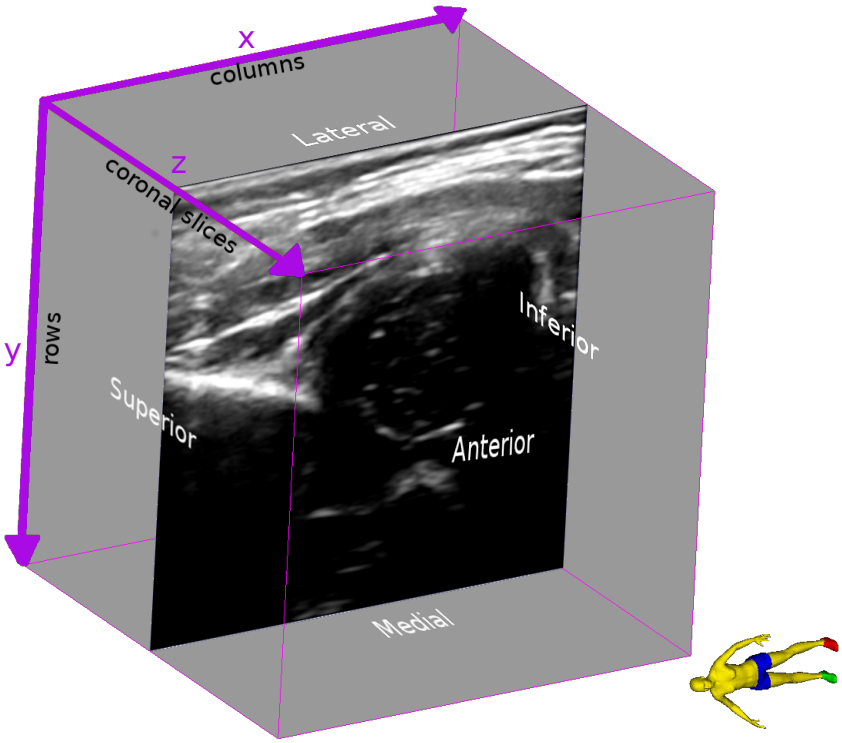


Figure 2.4: Coordinate system conventions used in this thesis.

Chapter 3

Comparative Evaluation of Hand-Engineered and Deep-Learned Features for Neonatal Hip Bone Segmentation in Ultrasound

In this chapter we address RQ-1 by implementing a popular data-driven model, namely U-Net [67], for pelvis bone surface segmentation and comparing it to the SOTA.

3.1 Methods

3.1.1 Hand-Crafted Features

We include CSPS [61–64] and SP [54] in our comparisons given their proven performance on neonatal hip US. Both methods tend to localize hip bone surfaces well but suffer from significant false positive responses at soft tissue, e.g. labrum and other irrelevant bone structures like the femur. To improve performance further we attempt to incorporate the spatial prior that the ilium and acetabulum are a continu-

ous bone structure that always appears as the most medial (or deepest with respect to the probe) and superior bone in the image.

To apply this spatial prior, we start with the observation that SP only detects one structure along each vertical scan line, which due to its high acoustic impedance, is likely to be bone. The hip bone is the most superior connected component of the SP segmentation. To find this region, we first define the set of k regions $CC = \{cc_1, ..., cc_i, ..., cc_k\}$ obtained by applying connected-component analysis to the SP binary segmentation mask, with 8-connectivity test in 2D and 26-connectivity in 3D. We further define the corresponding set of their centroids $C = \{c_1, ..., c_i, ..., c_k\}$, with $c_i = (x_i, y_i)$, and the corresponding set of x-components of the centroids $X = \{x_1, ..., x_i, ..., x_k\}$. We find the pelvis bone surface region cc_{xmin}, where the index $xmin = \arg\min(X)$. The final segmentation with SP is obtained by converting this connected region cc_{xmin} to a binary mask.

For CSPS, we first threshold the CSPS output volume to obtain a binary mask. Similarly, we apply connected-component analysis to obtain a set of connected regions. To find the pelvis bone surface from this set of regions, we leverage the segmentation previously obtained from SP of the pelvis bone surface, cc_{xmin}, as a Region-of-Interest (ROI), and use it to find the CSPS region with most overlap. We convert this to a binary mask, and to eliminate any soft tissue connected to the bone, we only keep the most medial (deepest) pixel along each scan line.

3.1.2 Deep-Learned Features

Architecture

We use the U-Net architecture [67] for our task of segmenting the ilium and acetabulum bone surfaces. We choose U-Net for its proven performance on medical image data and its ability to train on very few training samples. As in the original architecture [67], our implementation includes nine convolution blocks, five in the contracting path and four corresponding blocks in the expanding path. Each block is made up of six layers in the following order: conv3x3-batchnorm-ReLU-conv3x3-batchnorm-ReLU, in contrast to the original architecture which did not include batch normalization layers [29]. We use stride 1 for the 3x3 convolutions.

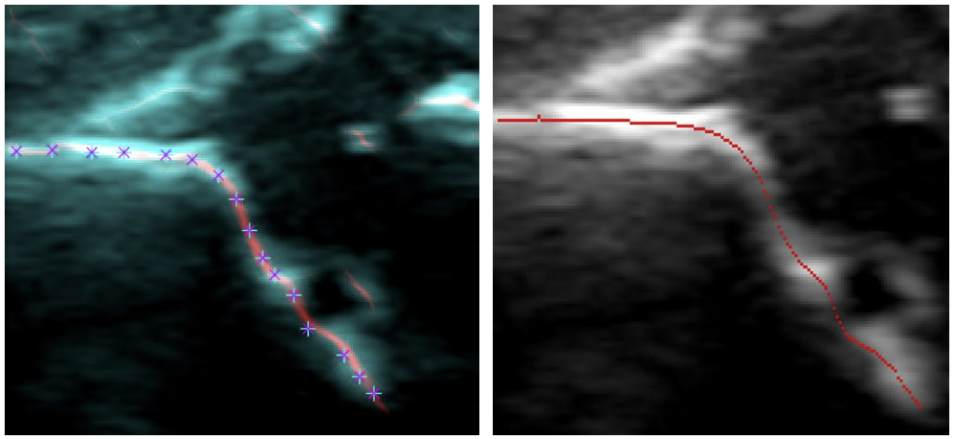


Figure 3.1: Example labeling procedure. **Left:** B-mode image with Structured Phase Symmetry overlaid in red, and user-defined points shown as asterisks. **Right:** contour fitted to user-defined points shown as a solid red line.

We use max pooling with 2x2 kernels and stride 2 in the contracting path, and corresponding transposed convolution layers in the expanding path. With this configuration, the receptive field of the convolution at the end of the contracting path is 140x140 pixels of the input image whose size is 250×250. The number of feature maps in the nine blocks is 64-128-256-512-1024-512-256-128-64. We explore training U-Net with two types of inputs: 1) B-mode only image data, and 2) a Multi-Channel (MC) input based on promising results from several recent papers [1, 74, 76] on bone segmentation that have shown improved accuracy of bone localization with this method. In our implementation, the multi-channel input includes the B-mode image, the corresponding SPS, and shadow confidence map [35] features in the R, G, B channels, respectively (see figure 3.2).

U-Net Training

To prepare the training data, we start with 231,384 coronal slices, obtained with the Ultrasonix probe from 59 neonate scans as potential training data. Approximately only 25% of these slices contained the anatomy of interest (ilium, acetabulum, femoral head), so we filter out all other slices with a recently proposed RNN scan

adequacy architecture [58]. We randomly select 500 such adequate slices, from which a trained user manually labelled the ilium and acetabulum bone contour. To aid with this manual training data labelling step, we overlaid the SPS feature on the B-mode image to help guide the user while allowing flexibility to deviate from SPS overlay should the user deem suitable (see figure 3.1). Further we allow the user to reject inadequate slices not detected by the RNN from the training set. In summary, we end up with 439 labelled samples in our training set from the Ultrasonix set, all of which contained the anatomy of interest. We intentionally did not include Clarius samples in the training set in order to test generalizability of U-Net on different domains. We subsequently dilate the manually traced contours as originally proposed by Villa to alleviate the class imbalance problem [74], the imbalance between the number of contour and background pixels, converting our bone contour to a ribbon-like structure. We train U-Net on both B-mode input only, as well as the multi-channel input. We select Dice loss, Adam optimizer, two-slice batch size, with learning rate of 0.0001 over 30 epochs, and resize the input images to 250×250 pixels.

3.1.3 Testing

We contrasted segmentation accuracy of five methods:

- Original CSPS (with naive thresholding)
- CSPS after applying the ROI prior as described in §3.1.1
- SP with ROI prior as described in §3.1.1
- U-Net with only B-mode input
- U-Net with multi-channel input data (R=B-Mode, G=SPS, B=Shadow Confidence Map)

We test these five methods on two datasets, a primary dataset of images that were acquired with the Ultrasonix 4DL14-5 3D-US probe, and a secondary dataset that was acquired with the Clarius L7 2D-US probe. The first test set was prepared from 880 3D-US volumes (from 25 neonate patients) using the Ultrasonix 4DL14-5/38 probe. Similar to the training data, we discarded inadequate slices with the

RNN approach [58] to randomly select a subset of adequate slices, on which the same user manually delineated the contours (see Figure 3.1). A final total of 103 labelled samples constituted this primary test set. We also prepared a secondary test set using data from the Clarius L7 probe, constituting 72 2D-US images from a different pool of 19 neonates.

Following segmentation using the contrasted methods, we performed simple post-processing to convert the output segmentation map output of U-Net to a crisp contour. Specifically, we threshold the probability map at 0.5, skeletonize, and prune the resulting binary mask to generate a contour.

To assess segmentation accuracy, many metrics have been proposed in the literature, but there is not a single standardized metric that encompasses all the information, so we include all metrics that may be applicable to our task. These can be categorized into two main groups: pixel-wise classification metrics, and open contour distance metrics. Classification metrics we reported include the following:

- Dice-Sorensen Coefficient (DSC), also known as the F1-score

$$DSC = \frac{2|R \cap P|}{|R| + |P|} = \frac{2TP}{2TP + FP + FN} \tag{3.1}$$

- Jaccard coefficient, also known as Intersection-Over-Union (IOU)

$$J = \frac{|R \cap P|}{|R \cup P|} = \frac{TP}{TP + FP + FN} \tag{3.2}$$

- Precision

$$Precision = \frac{TP}{TP + FP} \tag{3.3}$$

- Recall

$$Recall = \frac{TP}{TP + FN} \tag{3.4}$$

Where, TP, FP, and FN are true positive, false positive, and false negatives pixels, respectively; R are the set of pixels in the reference (ground truth) segmentation and P are the set of pixels in the predicted segmentation; $|\cdot|$ is the cardinality operator, which returns the number of elements in a set. Note that these pixel-wise metrics are conventionally used for measuring blob-shaped segmentations, and cannot

be used directly on thin, contour-like segmentations such as ours. To get around this, we dilated both reference and prediction contours equally as was previously proposed by others [74].

Distance metrics include:

- Vertical Root Mean Square Error (VRMSE), defined as the root mean squared distance between the reference and predicted contours at every scan line that contains both contours. More precisely, let $\boldsymbol{R} = r_{i,j} \in \mathbb{R}^{M\times 2}$ be the set of all points in the reference contour and $\boldsymbol{P} = p_{i,j} \in \mathbb{R}^{N\times 2}$ be the set of all points in the predicted contour. Define $\boldsymbol{R} \supseteq \boldsymbol{R}^o = r^o_{i,j} \in \mathbb{R}^{K\times 2}$ and $\boldsymbol{P} \supseteq \boldsymbol{P}^o = p^o_{i,j} \in \mathbb{R}^{K\times 2}$, the subsets of the contours where both reference and predicted contours exist. VRMSE is defined as,

$$VRMSE = \sqrt{\frac{\sum_i^K (r^o_{i,2} - p^o_{i,2})^2}{K}} \tag{3.5}$$

  Where $r^o_{i,2}$ is the y-component of the i-th element of $\boldsymbol{R}^o$, $p^o_{i,2}$ is the y-component of the i-th element of $\boldsymbol{P}^o$, and K is the number of rows in $\boldsymbol{R}^o$ and $\boldsymbol{P}^o$.

- Hausdorff Distance (HD). To compute HD, we first define the function $d(p,q)$ that computes the Euclidean Distance between two points p and q,

$$d(p,q) = \sqrt{(p_x - q_x)^2 + (p_y - q_y)^2} \tag{3.6}$$

  We define $\boldsymbol{A} = a_{ij} \in \mathbb{R}^{M\times N}$ as the matrix of Euclidean Distances from all points in $\boldsymbol{R}$ to all points in $\boldsymbol{P}$, calculated with $d()$. We compute the vector of minimum distances from each point in $\boldsymbol{R}$ to each point in $\boldsymbol{P}$ as the row-minima of $\boldsymbol{A}$,

$$d_{R2P} = \min_j a_{ij} \tag{3.7}$$

  And similarly we define the vector of minimum distances from each point in P to each point in R as the column-minima of $\boldsymbol{A}$,

$$d_{P2R} = \min_i a_{ij} \tag{3.8}$$

Table 3.1: Mean (and SD) segmentation accuracy of five methods we tested on the primary and secondary datasets. From left to right: 1) Shadow Peak with RoI spatial prior, 2) Confidence-Weighted Structured Phase Symmetry with naive thresholding, 3) Confidence-Weighted Structured Phase Symmetry with RoI spatial prior, 4) U-Net with B-mode input, 5) U-Net with multi-channel input. Best performers along each row are bolded.

Ultrasonix	*SP+RoI*	*CSPS*	*CSPS+RoI*	*U-Net*	*MC U-Net*
Jaccard	0.61 (0.13)	0.28 (0.14)	0.70 (0.16)	0.76 (0.10)	**0.77 (0.11)**
Dice-Sorensen	0.75 (0.12)	0.42 (0.16)	0.81 (0.14)	**0.86 (0.07)**	**0.86 (0.08)**
Precision	0.79 (0.12)	0.30 (0.15)	0.86 (0.11)	**0.89 (0.07)**	**0.89 (0.07)**
Recall	0.71 (0.14)	0.82 (0.11)	0.78 (0.17)	**0.85 (0.10)**	**0.85 (0.11)**
Hausdorff (mm)	4.41 (3.04)	21.89 (9.7)	3.06 (3.1)	**1.60 (1.67)**	1.91 (2.17)
VRMSE (mm)	0.35 (0.32)	5.45 (4.96)	0.37 (0.61)	0.21 (0.07)	**0.20 (0.07)**
Clarius	-	-	-	-	-
Jaccard	0.58 (0.08)	0.34 (0.09)	0.69 (0.14)	0.85 (0.07)	**0.86 (0.06)**
Dice-Sorensen	0.73 (0.06)	0.51 (0.10)	0.81 (0.10)	**0.92 (0.04)**	**0.92 (0.04)**
Precision	0.88 (0.04)	0.35 (0.09)	0.72 (0.15)	0.92 (0.07)	**0.94 (0.05)**
Recall	0.64 (0.09)	0.93 (0.06)	**0.95 (0.05)**	0.93 (0.03)	0.91 (0.05)
Hausdorff (mm)	5.79 (1.92)	25.68 (3.54)	5.65 (4.55)	2.34 (4.79)	**1.09 (0.90)**
VRMSE (mm)	0.33 (0.12)	2.42 (3.27)	0.33 (0.12)	0.22 (0.10)	**0.20 (0.07)**

Finally, HD is calculated as follows,

$$HD = \max(\max(d_{R2P}), \max(d_{P2R})) \tag{3.9}$$

3.2 Results and Discussion

Quantitative results for both the Ultrasonix and Clarius probe datasets are summarized in Tab. 3.1. Across all evaluation metrics, the B-mode U-Net and MC U-Net appeared to be virtually tied for best performance, and appeared to perform well consistently as evident by the reduced standard deviations. CSPS suffered from significant soft tissue false positives despite its use of shadow features, which explains its high recall but low precision rates, but we observe that the precision was much improved after applying the ROI, with a small drop in recall.

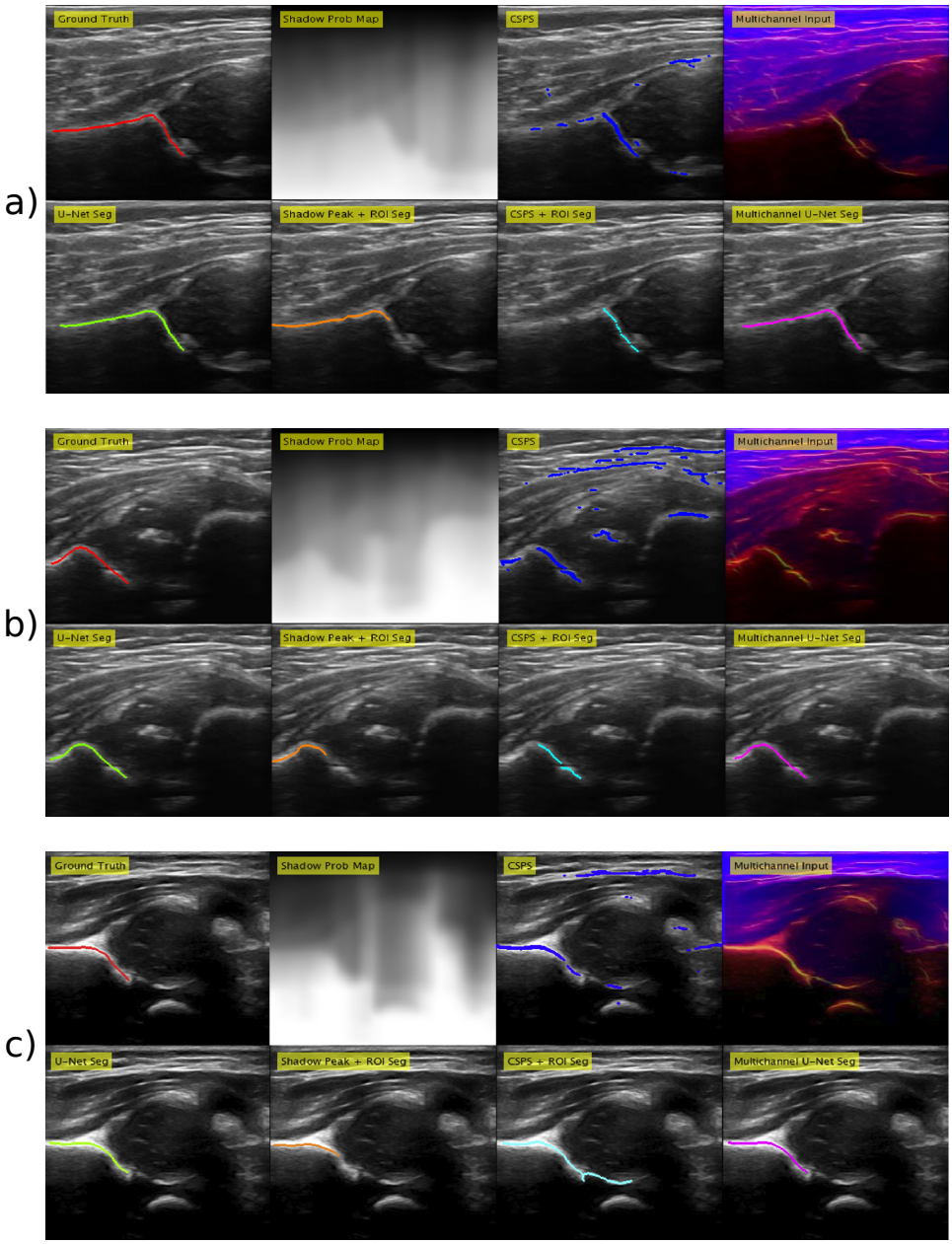


Figure 3.2: Example segmentation results. a,b) different segmentation techniques including SP, CSPS, and U-Net applied to Ultrasonix test data. c) the same techniques applied to Clarius test data.

To evaluate generalizability, we tested our model trained only on the Ultrasonix data on a secondary test set obtained with the Clarius probe. We saw a similar pattern on this secondary Clarius set, with U-Net and MC U-Net outperforming the other methods. Although not directly comparable, as the Clarius dataset was comprised of only 2D-US optimal coronal images of the infant hip, whereas the Ultrasonix dataset is 3D and contained coronal slices away from the optimal central slice, these results still provide some evidence that U-Net is likely capable of generalizing to image data from probes not included in the training set.

We show exemplar qualitative results on both Ultrasonix and Clarius data in Figure 3.2. To extract the α angle, it is crucial for the segmentation algorithm to accurately delineate the bony rim between the ilium and acetabulum, and enough of the ilium and acetabulum surfaces surrounding it, while simultaneously not capturing any false positive soft tissue or unrelated bone such as the femur. It is subsequently crucial to reduce outliers and carefully assess failure cases beyond mere aggregate quantitative measures comparisons such as those in Tab. 3.1. For CSPS, common failure cases included soft tissue false positives; completely missing the ilium when rotated at certain angles; as well as fragmented contour, as shown in Figure 3.2. Similarly, SP often missed the acetabulum due to weak shadow in that region (Figure 3.2). In contrast, U-Net rarely detected false positive soft tissue, and consistently and accurately segmented the full ilium and acetabulum contour. Errors were mainly due to slightly under- or over-segmenting the ilium or acetabulum at the superior and inferior extremities of the contour. Along each scan line, we observe that U-Net very accurately segmented the bone contours as is reflected in the negligible VRMSE errors observed on our clinical dataset.

In contrast to recent papers reporting improved results by fusing phase symmetry and shadow features within MC deep learning networks [1, 74, 76], we did not observe significant improvements by including these hand-crafted features to our input. However, an apparent improvement in HD indicates slight improvement on the secondary test set, and this is consistent with qualitative observations as we observed that MC U-Net is sometimes more robust to soft tissue false positives. Comparing to the closest literature, we observed on the primary test set improved mean HD of 1.60$\pm$1.67 mm compared to Hareendranathan's 2.1$\pm$0.9mm with the superpixel classification method [22]. Further we report much improved mean

DSC of 88% on our primary test set, compared to Zhang's [81] DSC of 39% for their proposed architecture, and 5% with their implementation of U-Net.

With regards to computational complexity, we had about 15 million parameters in our U-Net implementation. When tested on 250×250 2D coronal US slices, on a machine with Intel Core i-7 (4.0 GHz, 6 core) processor and NVIDIA Titan Xp GPU, we logged run times of 0.007s for Shadow Peak, 0.155s for CSPS, and 0.003s for U-Net.

3.3 Conclusions

We proposed a deep-learning-based approach for bone segmentation in neonatal hip ultrasound. We showed this method improved accuracy over state-of-the-art, feature-based techniques recently proposed in the literature for our task. Results on a secondary dataset show that U-Net is robust to domain shifts such as images from a probe that produces significantly different images, and that using a multi-channel input may improve robustness further. The main limitation of this experiment is using a CNN architecture that uses 2D convolution kernels, that do not incorporate potentially useful information from adjacent slices. We address this limitation in §4.1 with 3D-U-Net.

Chapter 4

Measuring Hip Dysplasia with 3-Dimensional Convolutional Neural Networks

In this chapter we address RQs-1,2, and 3. We address RQ-1 in §4.1, training *3D*-U-Net for pelvis bone surface segmentation, testing, comparing it to other methods. We address RQ-2 in §4.2, implemneting and training regression-based and segmentation-based CNNs for femoral head localization, testing, and comparing these models to other methods. We address RQ-3 in §4.3, proposing an algorithm for extracting key 3D-US DDH metrics from the segmented anatomy, and evaluate our algorithm's performance in a clinical study against the SOTA.

4.1 Pelvis Bone Surface Segmentation: Going 3D

In the previous chapter, we showed that CNNs can outperform the previously proposed CSPS for pelvis bone surface segmentation. However, U-Net [67] uses 2D convolutions, so can only process the input volume slice-by-slice. Recent architectures have been proposed that employ 3D convolutions to process the volume as a single input, making use of potentially important 3D information in adjacent slices. Most famously, these architectures include 3D-U-Net [6] and V-Net [46]. These architectures were concurrently published and are very similar to each other, with

the exception that V-Net additionally uses short-distance residual connections. In this section, we extend our work from the last chapter on pelvis bone surface segmentation by training 3D-U-Net and comparing its performance to U-Net [67] and CSPS [62].

4.1.1 Labeling

One of the main challenges in training CNNs is getting enough labelled data. This is especially a problem for getting segmentation mask annotations for volumetric data, as the time required for labeling only a single training sample (i.e. a full volume) is on the order of 10 to 100 times the time to label a single slice. In our case, labeling a single slice takes roughly 10s. There are usually around 50 slices with visible pelvis bone surface, so this translates to roughly 10 minutes per volume. This is the main reason we opted to use 2D U-Net in Ch.3 instead of directly using a 3D architecture such as 3D-U-Net. To reduce the time required for getting mask annotations for our task, we leverage the previously trained U-Net from Ch.3 for getting approximate mask annotations. We then manually fix any over- or under-segmentation in 3D Slicer [36] (see figure 4.1). We annotated the pelvis bone surface in a total of 116 volumes from a total of 29 participants, with 64 volumes for training and 52 for testing.

4.1.2 Training

We train 3D-U-Net to optimize Binary Cross Entropy (BCE) Loss,

$$L_{BCE}(y,\hat{y}) = -(y\log(\hat{y}) + (1-y)(\log 1-\hat{y})) \tag{4.1}$$

where y is target class for each pixel and $\hat{y}$ is the predicted class. We use batch size of 1 volume, and use 4-fold cross-validation to find a good learning rate of 0.001. We resize the input volume down to $100\times100\times100$ voxels. We train for 90 epochs, starting with an initial learning rate of 0.001 and reducing it by a factor of 0.2 at 30 and 60 epoch milestones. We use Adam optimizer [37]. Since we only have a relatively small training set of 64 training volumes, we apply the following random augmentations:

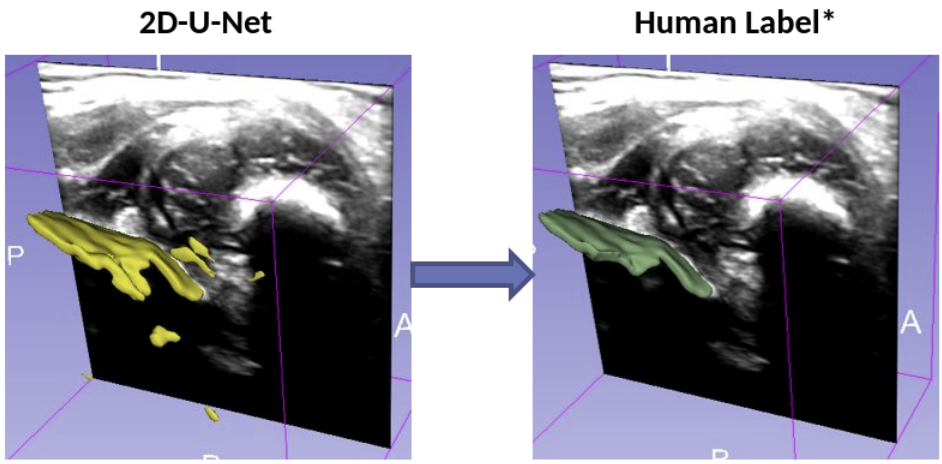


Figure 4.1: To train 3D-U-Net, we use U-Net predictions (left) from the previous chapter as a starting point and manually fix areas that are incorrectly segmented to get the "human" label (right).

- Non-uniform zooming by a factor in the range of [0.9, 1.1]
- Shifting along the x,y,z axes in the range of [-10, 10] pixels
- Rotating around the x,y,z axes in the range of [-5, 5] degrees
- Flips in the medial/lateral direction (z-axis) with 0.5 probability
- Elastic deformation [72] with probability 0.5, and σ in the range [2,4]
- Gamma contrast correction with γ in the range of [0.2,2]

$$I_{out} = I_{in}^{\gamma} \tag{4.2}$$

4.1.3 Testing

To evaluate the performance of the different methods proposed, we assess agreement with a labelled test set of 52 volumes from 13 participants. In this section, we compare the following methods:

- 3D-U-Net [6]

- U-Net [67] that was proposed in Ch.3
- CSPS as originally proposed by Quader [62], with naive thresholding
- CSPS-DDH, which is the CSPS-based segmentation that Quader used to compute the DDH metrics such as α_{3D} [61]. This included some post-processing such as cropping with an ROI and ray-casting.

Note that in this comparison we do not include methods previously evaluated in Ch.3 including SP [54], MC U-Net, and CSPS with connected-component analysis and anatomical prior. This is because, based on the evidence we saw in Ch.3, U-Net performed the best and we do not see an added benefit of further investigating the other methods.

As in §3.3, we use both classification and distance metrics to assess accuracy of segmentation. Classification metrics used in this chapter include the following:

- Dice-Sorensen Coefficient (DSC), also known as the F1-score (see Eq.3.1)
- Jaccard coefficient, also known as IOU (see Eq.3.2)
- Precision (see Eq.3.3)
- Recall (see Eq.3.4)

Note that all of these metrics are conventionally used for measuring blob-shaped segmentations, and cannot be used directly on thin, contour-like segmentations such as ours. To get around this, we dilated both reference and prediction contours equally with a $5\times5\times5$ voxel cubic structuring element, as was previously proposed by others [74].

For measuring distance between the reference and predicted contours we use:

- Mean Euclidean Distance (MED), and this can be computed from the reference surface to the predicted surface, prediction to reference, or bidirectionally as the maximum of these. To calculate the MED from reference to prediction, let $\boldsymbol{R} = r_{i,j} \in \mathbb{R}^{M\times3}$ be the set of all points in the reference surface and $\boldsymbol{P} = p_{i,j} \in \mathbb{R}^{N\times3}$ be the set of all points in the predicted surface.

Further we define the function $d(p,q)$ that computes the Euclidean Distance between two points p and q,

$$d(p,q) = \sqrt{(p_x - q_x)^2 + (p_y - q_y)^2 + (p_z - q_z)^2} \tag{4.3}$$

We define $\boldsymbol{A} = a_{ij} \in \mathbb{R}^{M \times N}$ as the matrix of Euclidean Distances from all points in $\boldsymbol{R}$ to all points in $\boldsymbol{P}$, calculated with $d()$. We compute the vector of minimum distances from each point in $\boldsymbol{R}$ to each point in $\boldsymbol{P}$ as the row-minima of $\boldsymbol{A}$,

$$d_{R2P} = \min_j a_{ij} \tag{4.4}$$

And similarly we define the vector of minimum distances from each point in $\boldsymbol{P}$ to each point in $\boldsymbol{R}$ as the column-minima of $\boldsymbol{A}$,

$$d_{P2R} = \min_i a_{ij} \tag{4.5}$$

Finally, we compute the MEDs from $\boldsymbol{R}$ to $\boldsymbol{P}$, $\boldsymbol{P}$ to $\boldsymbol{R}$, and bidirectionally as,

$$MED_{R2P} = \frac{\sum d_{R2P}}{N} \tag{4.6}$$

$$MED_{P2R} = \frac{\sum d_{R2P}}{M} \tag{4.7}$$

$$MED_{max} = \max(MED_{R2P}, MED_{P2R}) \tag{4.8}$$

- Hausdorff Distance (HD), and similarly this is computed from reference to prediction, prediction to reference, or bidirectionally as the maximum of these.

$$HD_{R2P} = \max(d_{R2P}) \tag{4.9}$$

$$HD_{P2R} = \max(d_{P2R}) \tag{4.10}$$

$$HD_{max} = \max(HD_{R2P}, HD_{P2R}) \tag{4.11}$$

In addition, our lab has recently proposed two new metrics that aim to ad-

dress some of the shortcomings of the aforementioned metrics, specifically for the task of bone segmentation in US, by combining classification and distance metrics into a unified measure, that does not rely on dilation. This unified measure is the Coverage Agreement Index (CAI) and is computed from the Coverage Distance Index (CDI) and Root Mean Square (RMS) distance as follows.

Given a 3D binary segmentation $\boldsymbol{B} = b_{i,j,k} \in \mathbb{Z}_2^{M \times N \times K}$, we compute the scan-line 2D binary projection $\boldsymbol{B}^{BP} = b_{j,k} \in \mathbb{Z}_2^{N \times K}$ as follows:

$$b_{j,k}^{BP} = \begin{cases} 1 & \text{if } \sum_i b_{j,k} > 0 \\ 0 & \text{otherwise} \end{cases} \tag{4.12}$$

Given the reference binary segmentation volume $\boldsymbol{R}$ and corresponding scan-line projection $\boldsymbol{R}^{BP}$, and the predicted binary segmentation volume $\boldsymbol{P}$ and corresponding scan-line projection $\boldsymbol{P}^{BP}$, the CAI is the DSC between the 2D binary projections and is computed as follows:

$$CAI = \frac{2|\boldsymbol{R}^{BP} \cap \boldsymbol{P}^{BP}|}{|\boldsymbol{R}^{BP}| + |\boldsymbol{P}^{BP}|} \tag{4.13}$$

Further, we compute the RMS Euclidean Distance error. For each image, it is the RMS of the Euclidean Distances from each point on the predicted bone surface to the nearest point on the reference bone surface.

$$RMS_{P2R} = \sqrt{\frac{\sum d_{P2R}}{M}}^2 \tag{4.14}$$

Where d_{P2R} was defined in equation 4.5.

Finally CDI is computed as:

$$CDI = \frac{CAI}{1 + RMS_{P2R}^2} \tag{4.15}$$

4.1.4 Results and Discussion

Testing results for the four contrasted methods are shown as boxplots in figures 4.3, 4.4, and 4.5, and a summary is provided in table B.1. As well, we show qualitative

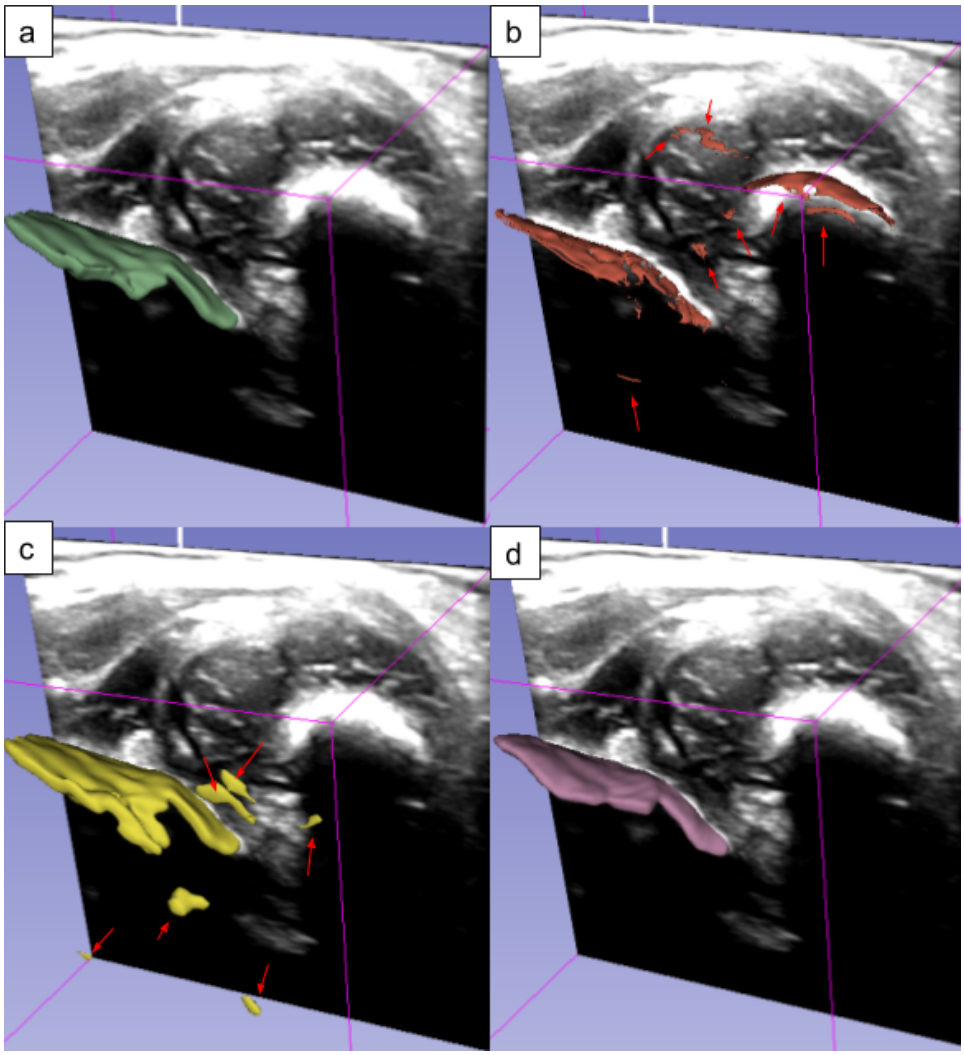
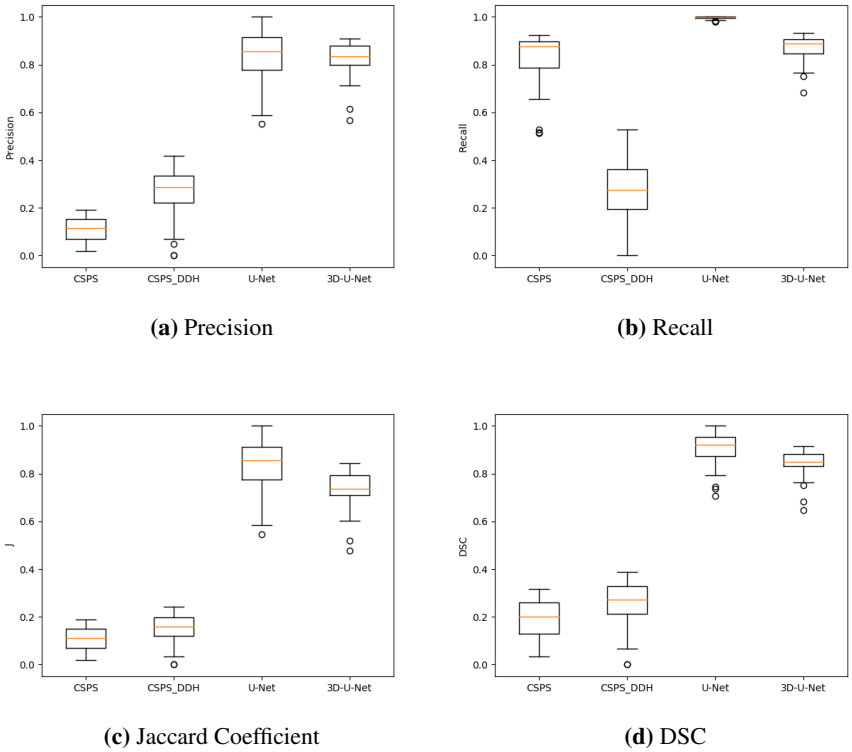


Figure 4.2: Visualizing pelvis bone surface segmentation with the contrasted methods. a) human label; b) CSPS; c) U-Net; d) 3D-U-Net. Red arrows point to areas that are over-segmented (false positives).



(a) Precision

(b) Recall

(c) Jaccard Coefficient

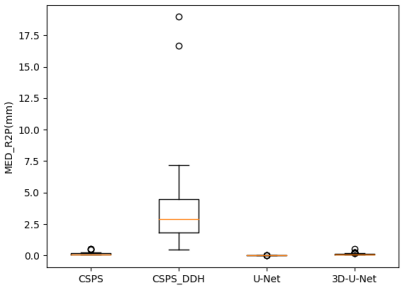
(d) DSC

Figure 4.3: Pixel-wise classification evaluation for pelvis bone surface segmentation.

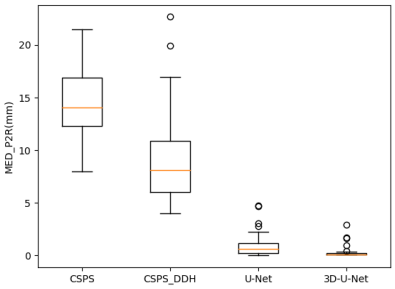
visual results in figure 4.2.

We perform statistical analyses as follows. We first test the null hypothesis that all four methods produce equivalent segmentations with a one-way Analysis of Variance (ANOVA) (see results in table B.1). Considering the conservative Bonferroni criterion for multiple tests [4], our p-value threshold for statistical significance is reduced from 0.05 down to 0.004. Even with this conservative criterion, we reject the null hypothesis across all 12 reported metrics. Further we apply post hoc t-tests between methods for all the reported metrics (see t-test results in Appendix B). From this, we make the following observations about the three key metrics:

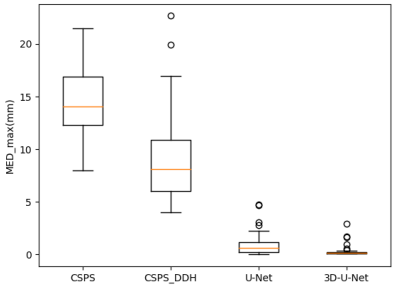
- DSC: U-Net and 3D-U-Net outperform CSPS. U-Net slightly outperforms



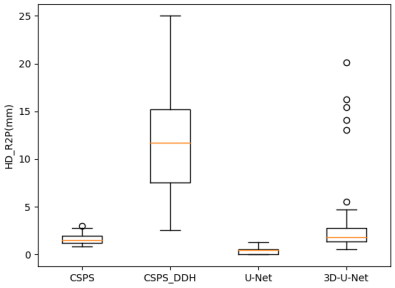
(a) MED_{R2P}



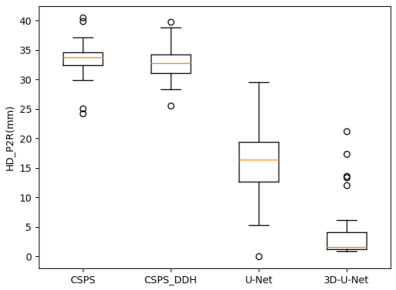
(b) MED_{P2R}



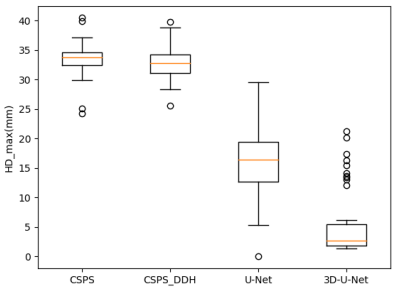
(c) MED_{max}



(d) HD_{R2P}

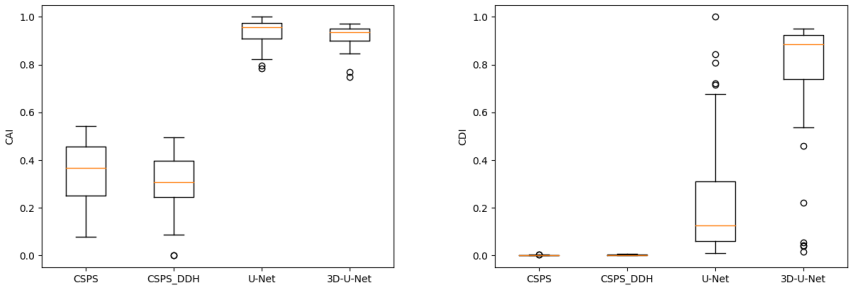


(e) HD_{P2R}



(f) HD_{max}

Figure 4.4: Contour distance evaluation metrics for pelvis bone surface segmentation.



(a) CAI **(b)** CDI

Figure 4.5: Combined evaluation metrics for pelvis bone surface segmentation.

3D-U-Net.

- HD_{max} and MED_{max}: 3D-U-Net outperforms all the other methods, including U-Net.
- CAI: 3D-U-Net far outperforms the other methods.

Based on these observations, as well as our qualitative visualizations of the results, we conclude that 3D-U-Net is our best option to use in the final algorithm for extracting the DDH metrics. The fact that U-Net slightly outperforms 3D-U-Net in the classification metrics and CDI is probably explained by the fact that our reference (ground truth) labels are initially based on the U-Net predictions. DSC of 85% for 3D-U-Net is still an objectively good score, and taking all the other metrics into consideration, we conclude that 3D-U-Net outperforms U-Net and far outperforms CSPS.

4.1.5 Conclusions

We proposed 3D-U-Net to segment the pelvis bone surface.3D-U-Net more accurately segmented the pelvis bone surface compared to CSPS, the SOTA method previously proposed by Quader [61–64] to segment the pelvis bone surface. Compared to the U-Net model proposed in chapter 3, 3D-U-Net captured similar bone

surface as U-Net, but with fewer false positive detached islands from the main bone surface.

4.2 Locating the Femoral Head

The second important landmark to delineate is the femoral head. This is especially true for extracting metrics such as FHC_{3D}. However, the femoral head presents its own unique challenges. Being almost completely cartilaginous in neonates, and therefore hypoechoic, the femoral head has weakly defined boundaries and appears as a dark, approximately semi-spherical shape with some speckle. It is bounded medially by hyperechoic pulvinar fat in the acetabulum, laterally by soft tissue including muscles and ligaments, and superiorly by the labrum and hypoechoic hyaline cartilage lining the bony rim (junction between the ilium and the acetabulum).

In this section, we explore two potential approaches for locating the femoral head:

1. Direct *regression* of sphere parameters (§4.2.2): We assume that the femoral head is a perfect sphere, although only part of the sphere is visible in an US volume. We use 3D-CNNs to directly regress the four sphere parameters, including the the center coordinates (c_x, c_y, c_z) and the radius.

2. *Segmentation* (§4.2.3): We do not make assumptions about the shape, and only attempt to segment the visible part of the femoral head. We again use 3D-U-Net [6] for this task.

4.2.1 Labeling

Similarly to the previous chapters, we need volumes with the femoral head annotated to train the proposed CNNs. As the femoral head does not have well-defined boundaries, we do not attempt a direct pixel-wise annotation as is customarily done for segmentation datasets. We found that determining which pixels are femoral head and which are background to be difficult. Additionally, frame-by-frame pixel-wise labeling is extremely time-consuming and inefficient. Instead, we make use of the knowledge that the femoral head is approximately spherical, and label it by

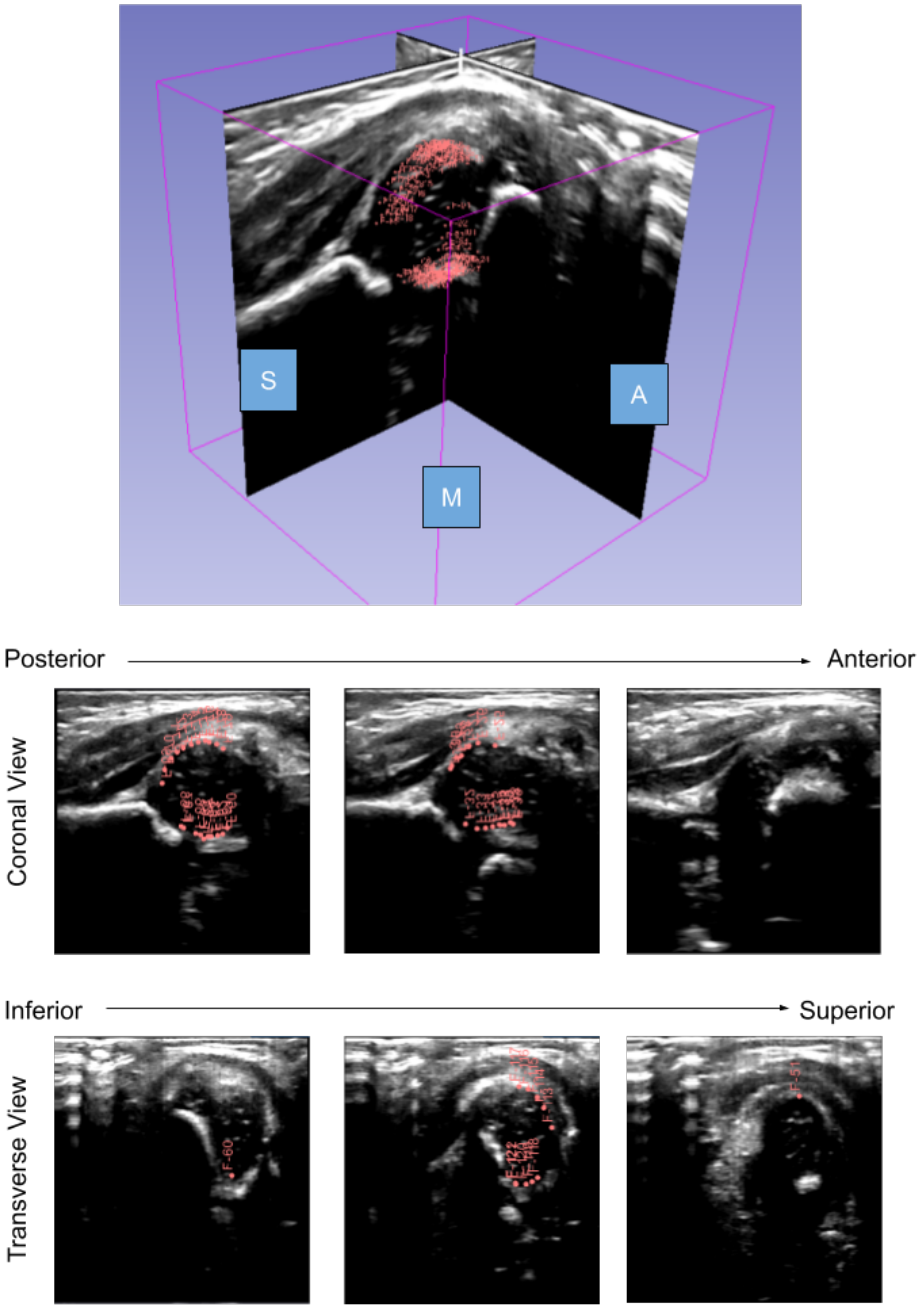


Figure 4.6: Femoral head keypoint placement in 3D Slicer.

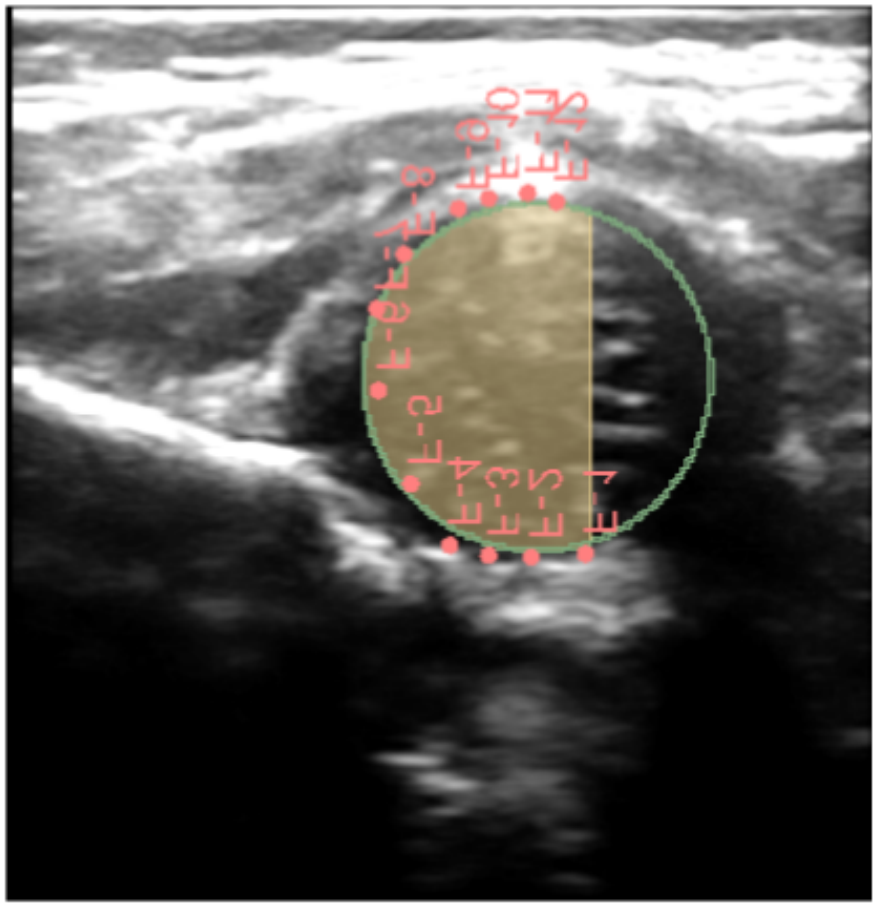


Figure 4.7: Fitting a sphere to the key edge points. We show the full sphere in green and the cropped sphere in yellow.

selecting only a few keypoints at its edges in some coronal and transverse frames. We demonstrate this process in figure 4.6. After selecting these key edge points, we use them to prepare two kinds of labels (see figure 4.7):

1. **Full sphere label**: We fit a sphere to these points with a least-squares method [33], generating a 4-element vector that includes the three center coordinates and the radius $[c_x, c_y, c_z, r]$

2. **Semi-sphere label**: From this fitted sphere, we generate a semi-spherical

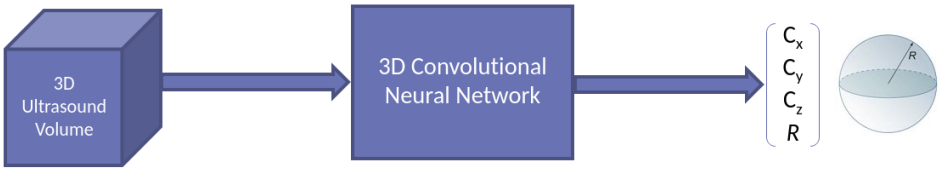


Figure 4.8: Conceptual illustration of 3D convolutional neural networks used for direct regression of sphere parameters.

binary segmentation mask of the visible part of the femoral head (see figure 4.7). First, any pixels within the sphere are labeled as 1 and any pixels outside are labeled as 0. Further, we define a bounding box B whose boundaries are defined by the most extreme keypoints. We set any points outside of B to 0, ensuring that only the clearly bounded, visible parts of the femoral head are segmented.

Our train set for this task contains 52 volumes from 13 participants, and the test set constains 48 volumes from 12 participants.

4.2.2 Direct Regression

Architecture Choice

Considering the weakly-defined boundaries of the femoral head in 3D-US of the neonatal hip, our initial guess was that formulating this problem as a *segmentation* problem would not be effective. Instead, we propose to directly *regress* the centre and radius parameters of the sphere-of-best-fit as shown in figure 4.8. Although CNNs have been mainly used for classification, they can be used just as effectively for regression. Sphere regression is by definition a 3D task, so we choose to use 3D models that use 3D convolutions, and have been shown to be effective for 3D tasks such as video classification [18–20]. 3D versions of modern architectures have been recently proposed including ResNet [26], whose residual connections were shown to be improve accuracy and efficiency over its predecessors [5]; as well as newer derivatives of ResNet including DenseNet [28], WideResNet [80],

and ResNext [79]. Volumetric data and architectures require significantly increased GPU memory and time for training. So, due to time and hardware limitations, we choose to limit our experiments to 3D-ResNet-50 and 3D-DenseNet-121 proposed in [18–20].

Training

We set up our models as shown in figure 4.8. We resize the input volume to $100 \times 100 \times 100$ voxels to meet memory constraints of our system. To identify a good model and hyper-parameters for our task, we use 4-fold cross-validation and experiment with architectures including DenseNet-121 and ResNet-50, augmentation options, and learning rates. We choose Mean Squared Error (MSE) loss as the objective function to train our networks for regression.

$$L_{MSE} = ||y - \hat{y}|| \tag{4.16}$$

Where y is the target label and $\hat{y}$ is the prediction. Based on results from the cross-validation, we finally choose 3D-ResNet-50, trained with batch-size of 3 volumes, for 210 epochs, with an initial learning rate of 0.001, reduced by a factor of 0.2 at the 70 epoch and 140 epoch milestones. We apply the following augmentation in training:

- Non-uniform zooming by a factor in the range of [0.9, 1.1]
- Shifting along the x,y,z axes in the range of [-10, 10] pixels
- Rotating around the x,y,z axes in the range of [-5, 5] degrees
- Flips in the medial/lateral direction (z-axis) with 0.5 probability
- Gamma contrast correction with γ in the range of [0.2, 2]

4.2.3 Segmentation

Considering our previous successes with U-Net and 3D-U-Net, as well as overwhelming evidence from the literature on the success of these architectures [30],

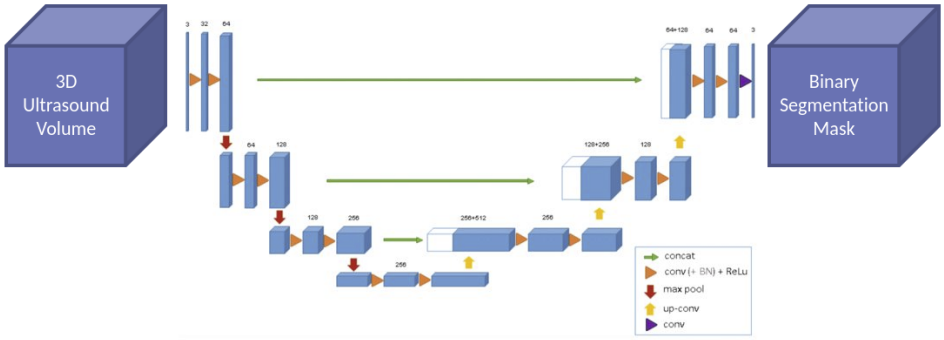


Figure 4.9: Conceptual illustration of 3D-U-Net used for segmenting the femoral head.

we evaluate the performance of 3D-U-Net for this task (figure 4.9), despite our initial doubts about the challenging task of segmenting such ill-defined anatomy. We use BCE loss (Eq. 4.1) as the objective function for this task. We use batch size of 1 volume, and train for 30 epochs with an initial learning rate of 0.0001, reduced by a factor of 0.2 at 10 and 20 epoch milestones. We use the same augmentations as in §4.2.2.

4.2.4 Testing

On a test set of 48 volumes from 12 participants, we evaluate the performance of our proposed ResNet-50 for direct sphere regression and 3D-U-Net for segmentation, against the previously proposed SOTA RFC by Quader [64]. Metrics we used for assessing performance can be divided into two categories: classification metrics and distance metrics.

Classification metrics include:

- Precision (Eq. 3.3)
- Recall (Eq. 3.4)
- Jaccard coefficient, also known as IOU (Eq. 3.2)
- DSC (Eq. 3.1)

Distance metrics include:

- Centre Absolute Error (CAE)s between the true and predicted spheres' centers along all three x, y, and z axes. Note that the 3D-U-Net segmentation ground truth label as well as its predicted output are not spherical, but instead are semi-spherical cropped by the bounding box B. So for the segmentation binary mask produced by 3D-U-Net, the center is computed as the center-of-mass of the segmented region. The errors are computed as follow:

$$CAE_x = |c_x - \hat{c}_x| \tag{4.17}$$

$$CAE_y = |c_y - \hat{c}_y| \tag{4.18}$$

$$CAE_z = |c_z - \hat{c}_z| \tag{4.19}$$

  Where c is the reference (ground truth) centre and $\hat{c}$ is the predicted centre.

- RAE between the true and predicted spheres' radii. Again, the 3D-U-Net prediction is not spherical, so the radius is computed as the difference between the most medial and lateral points of the segmented region.

$$RAE = |r - \hat{r}| \tag{4.20}$$

  Where r is the reference (ground truth) radius and $\hat{r}$ is the predicted radius.

- Centre Euclidean Distance (CED) between the predicted center and human label, computed as follows:

$$CED = \sqrt{(c_x - \hat{c}_x)^2 + (c_y - \hat{c}_y)^2 + (c_z - \hat{c}_z)^2} \tag{4.21}$$

4.2.5 Results and Discussion

We report quantitative testing results as boxplots in figures 4.13 and 4.14, as well a summary in table C.1. Qualitative, visual examples are shown in figures 4.10, 4.11, and 4.12.

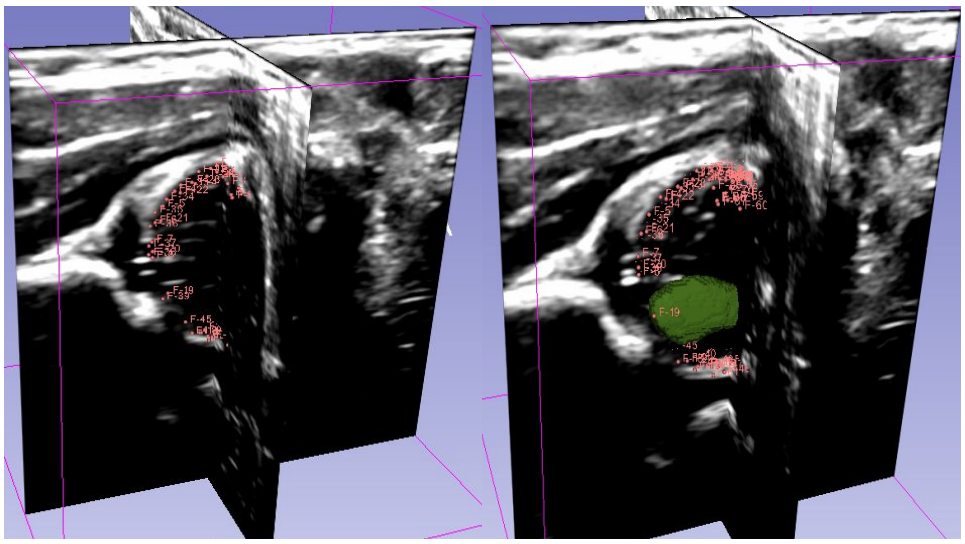


Figure 4.10: Visualizing Quader's [61, 64] RFC prediction of the femoral head. Left: hip ultrasound with human-labelled keypoints of the femoral head in red. Right: binary segmentation mask output of the RFC in green.

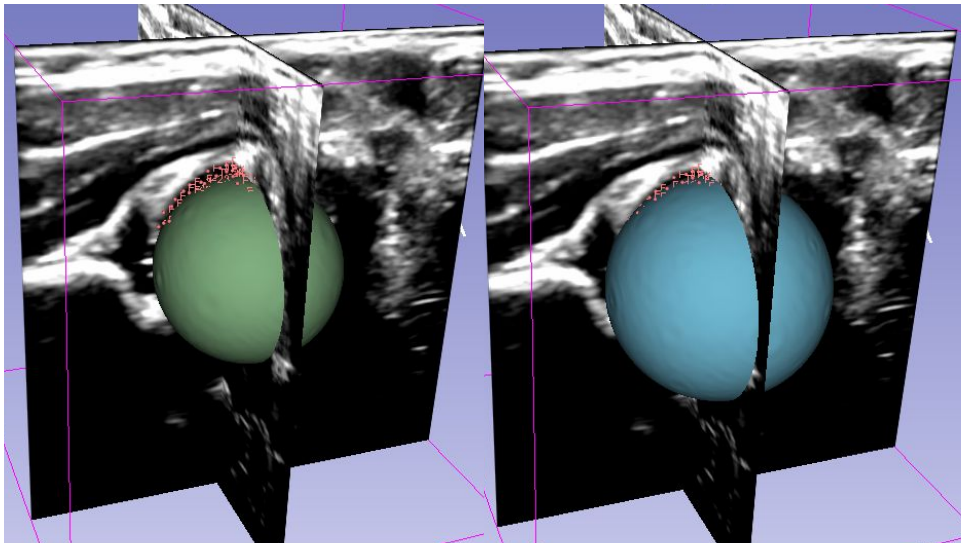


Figure 4.11: Visualizing output of 3D-ResNet-50 direct regression model for femoral head localization. Left: hip ultrasound with human-labelled keypoints of the femoral head in red, and best fitting sphere in green. Right: 3D-ResNet-50 predicted sphere in blue.

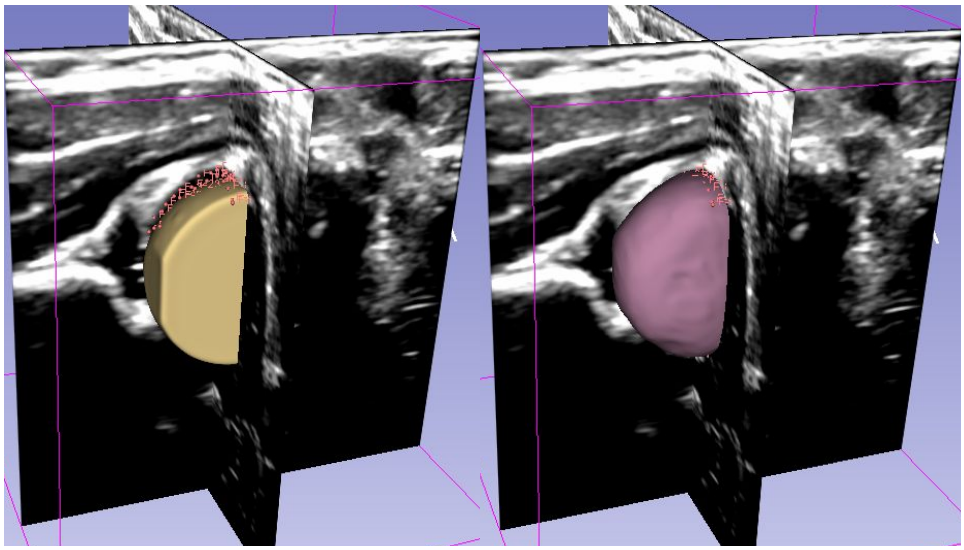
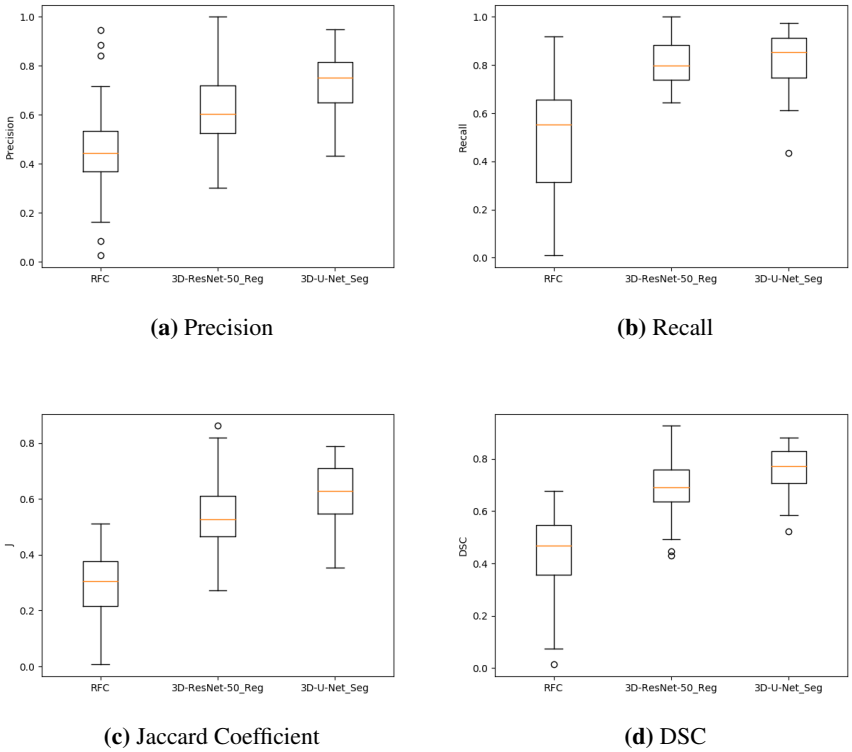


Figure 4.12: Visualizing output of 3D-U-Net segmentation model for femoral head localization. Left: hip ultrasound with human-labelled keypoints of the femoral head in red, and best fitting, cropped sphere in yellow. Right: 3D-U-Net segmentation binary mask prediction in pink.

Similar to the previous section we apply the following statistical analyses. We apply a one-way ANOVA to test the null hypothesis that all 3 methods are equal (see results of ANOVA in table C.1). Considering the Bonferroni correction for multiple comparisons [4] (in our case 9 comparisons), our p-value threshold for statistical significance is reduced from 0.05 down to 0.006. With this conservative threshold, we can reject this null hypothesis for all the metrics except for the CAE_x, for which a p-value of 0.009 suggests we cannot reject the null hypothesis for this metric. Further, we apply post hoc t-tests to compare the different methods (see tables in Appendix C). We make the following observations about the key metrics:

- DSC: both of our proposed models, 3D-U-Net (segmentation) and 3D-ResNet-50 (regression) outperform the RFC, and 3D-U-Net performs the best.
- CED: 3D-U-Net predicts the center against its ground truth label most accurately. 3D-ResNet-50 and the RFC are virtually tied.
- RAE: 3D-ResNet-50 and 3D-U-Net outperform the RFC, and 3D-U-Net pre-



(a) Precision

(b) Recall

(c) Jaccard Coefficient

(d) DSC

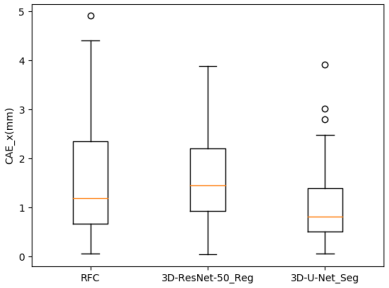
Figure 4.13: Pixel-wise classification-based evaluation for femoral head localization.

dicts the radius most accurately.

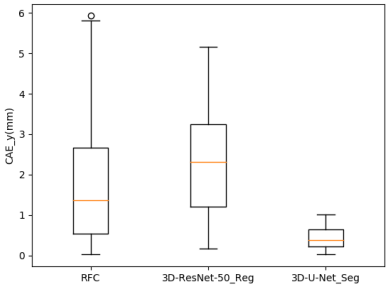
Based on these observations and our qualitative observations of the visual segmentations, we conclude that 3D-U-Net is the best of the three methods presented, so we choose to use it in our final pipeline for DDH metrics extraction.

4.2.6 Conclusions

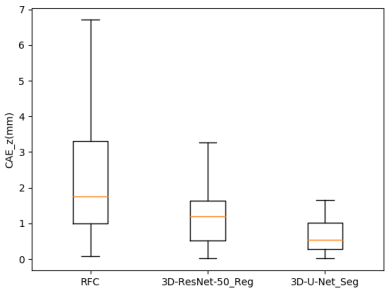
We proposed two new methods based on 3D-CNNs for locating the femoral head in neonatal hip 3D-US. We directly compared the performance of our methods to each other, as well as to the SOTA RFC proposed by Quader [61, 64], the only other fully automatic method proposed for our task to the best of our knowledge.



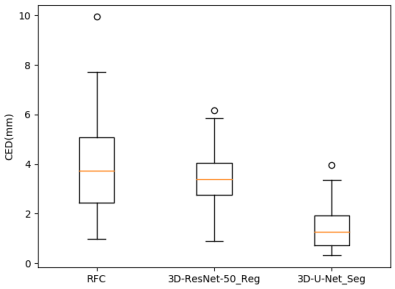
(a) CAE_x



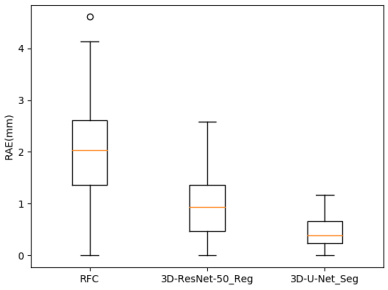
(b) CAE_y



(c) CAE_z



(d) CED



(e) RAE

Figure 4.14: Distance-based evaluation metrics for femoral head localization.

We found that the regression-based, 3D-ResNet-50 model locates the femoral head more accurately than the RFC in all the proposed metrics, except for the CAE_x. The segmentation-based, 3D-U-Net model locates the femoral locates more accurately than 3D-ResNet-50 and the RFC, so we choose to use 3D-U-Net in our final metrics pipeline for extracting dysplasia metrics as described in the next section.

4.3 Extracting Dysplasia Metrics

In this section, we describe our algorithms for extracting α_{3D}, FHC_{3D}, and OCR DDH diagnostic metrics from neonatal hip 3D-US, using the improved segmentation techniques described in the previous chapters. In addition, we present a clinical study to evaluate the performance of our algorithms against the SOTA.

4.3.1 Choosing DDH Metrics

Many US-based diagnostic metrics have been previously presented in the literature for DDH [66]. For consistency with standard clinical practice, and for direct comparison to the SOTA algorithms, we choose to focus only on the following three metrics:

- $\boldsymbol{\alpha}_{3D}$: a 3D metric that was first proposed by Quader [63], analogous to the widely clinically-used α angle first proposed by Graf [12]. This metric is defined as the angle between the normals to the fitted planar surfaces of the ilium and acetabulum. DDH severity increases with decreased α_{3D} angle.

- **FHC**$_{3D}$: another 3D metric that was first proposed by Quader [64], analogous to the widely used FHC metric originally proposed by Morin [47]. This metric captures additional information not captured by α_{3D}, and can potentially be used for quantitative dynamic assessment as proposed by Paserin [59]. This metric is defined as the ratio of the femoral head volume medial to the plane of the ilium vs. the total femoral head volume. DDH severity increases with decreased FHC_{3D}.

- **OCR**: another 3D metric that was initially proposed by Zonoobi and Hareendranathan [82]. This metric is relatively new and is not traditionally used in standard clinical practice. We choose to report it as it may provide additional

information about bony rim rounding that is not necessarily captured by the α_{3D} angle. This metric is defined as the radius of the largest sphere that can be fitted under the bony rim (junction between ilium and acetabulum, also called "apex line"). DDH severity increases with increased OCR.

4.3.2 Algorithm for Extracting the Metrics

We propose new methods for extracting these 3 metrics from the 3D-U-Net pelvis bone and femoral head segmentations, that builds on ideas from the previously proposed algorithms by Quader [61] and Hareendranathan [82]. The algorithm is illustrated in figure 4.15.

Getting OCR

Starting with the pelvis bone surface segmentation, we do the following to extract the OCR:

1. Apply connected-component analysis to remove any small detached islands from the main surface of interest, keeping only the largest component.

2. Skeletonize to convert the thick segmentation to a thin (one-pixel-wide) surface binary segmentation.

3. Convert this binary segmentation to a point cloud $\mathbf{PC_P}$.

4. Fit a polynomial surface $S_P(x,z)$ of the 2nd order along the z-axis and 3rd order along the x-axis to $\mathbf{PC_P}$. This allows us to compute the surface Gaussian Curvature K in the next step, as the point cloud surface is otherwise too noisy for this calculation.

5. Compute the Gaussian Curvature $K(x,z)$ of the polynomial surface $S_P(x,z)$ as described by Zonoobi and Hareendranathan [82].

6. Find the coordinates (x_K, z_K) of the point of maximum Gaussian Curvature, $R = (x_K, z_K, y_K)$, on the polynomial surface.

$$(x_K, z_K) = \underset{(x,z)}{\operatorname{argmax}} \, |K(x,z)| \tag{4.22}$$

7. Compute OCR as the reciprocal of the First Principal Curvature K_1 at the point of maximum K:

$$OCR = \frac{1}{|K_1(x_K, z_K)|} \tag{4.23}$$

Getting α_{3D}

Building on the steps for calculating OCR, we do the following to extract α_{3D}:

1. Similar to the pelvis bone segmentation, we convert the femoral head binary segmentation to a point cloud $\mathbf{PC_F} = f_{i,j} \in \mathbb{R}^{N\times 3}$.

2. Find the center-of-mass C of the femoral head as

$$C = \frac{\sum_i f_i}{N} \tag{4.24}$$

3. Define the sphere O with center C and radius $\|C - R\|$, which will be used to separate the ilium and acetabulum point clouds

4. Assign all the points in $\mathbf{PC_P}$ and outside O to the ilium point cloud $\mathbf{PC_I}$

5. Assign all the points in $\mathbf{PC_P}$ and inside O to the acetabulum point cloud $\mathbf{PC_A}$

6. Fit planes A and I to the $\mathbf{PC_A}$ and $\mathbf{PC_I}$ point clouds, respectively, with least-squares plane-fitting

7. Compute α_{3D} as the angle between the unit normal vector to the iliac plane n_I and the unit normal vector to the acetabulum plane n_A:

$$\alpha_{3D} = \cos^{-1} \frac{n_A \cdot n_I}{|n_A||n_I|} \tag{4.25}$$

Getting FHC_{3D}

Further building on the previous steps, we simply compute FHC_{3D} as,

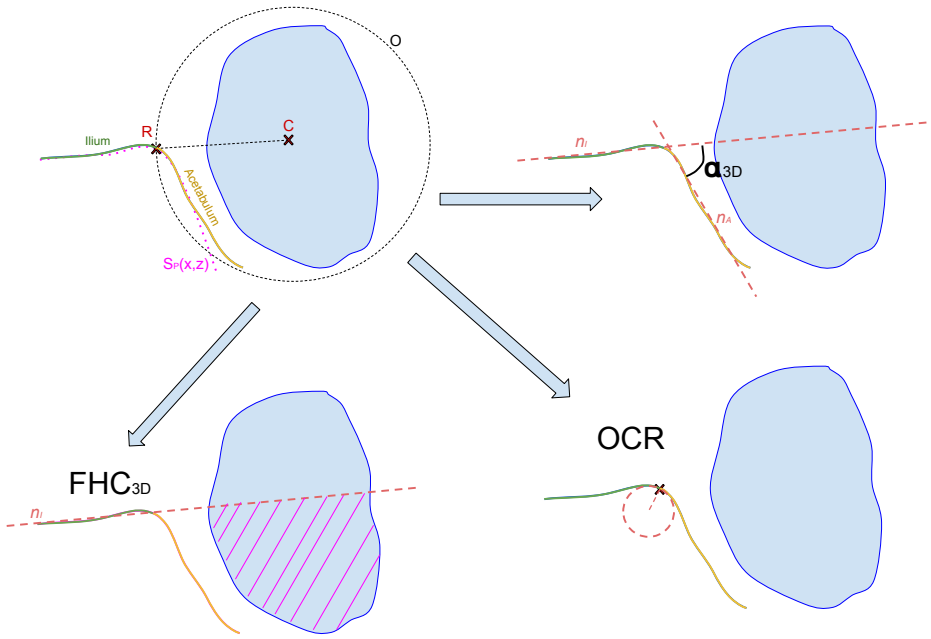


Figure 4.15: Conceptual illustration of proposed pipeline for extracting α_{3D}, FHC_{3D}, and OCR from the segmented pelvis bone surface and femoral head. Note that measurement is done in 3D, but concept is simplified to 2D for illustration purposes only.

$$FHC_{3D} = \frac{\sum B_F^M}{\sum B_F} \tag{4.26}$$

Where B_F is the femoral head binary segmentation mask, and B_F^M is the binary segmentation mask containing only the femoral head portion medial to the plane of the ilium I.

4.3.3 Clinical Study

We conduct a clinical study to evaluate the performance of our proposed algorithms against the SOTA methods. Ideally, we would assess *accuracy* of our methods against a clinical *gold standard*, but such a measure does not exist because 2D-US, the current clinical standard, is highly variable as previously explained so cannot be considered a gold standard. Accuracy could be assessed with a longitu-

dinal study that tracks patient outcomes, but this type of study requires long-term tracking on the order of years and is beyond the scope of this thesis. Therefore, we only compare the *reliability* of our methods to the state-of-the-art techniques. Additionally, we perform a comparative visual comparison of the different methods when large discrepancies with previous methods are observed.

Reliability

To assess reliability we use 483 sweeps from 34 participants as was described in Ch.2. Our goal is to simulate the real clinical scenario as close as possible, so we design our study to assess inter-exam test-retest reproducibility. In our study, an expert US technician images both hips of every participant. The same technician scans each hip twice (i.e. 2 exams), removing and replacing the probe in between exams. Within an exam, the probe is held still while the transducer is swept anteriorly-posteriorly 4 times. In summary, each participant is imaged 16 times, with each hip imaged twice (2 exams), and each exam containing 4 sweeps. See Figure 2.2 for an illustration. We note that the same technician does both exams, so we can only compute the *intra*-sonographer reproducibility and not the *inter*-sonographer reproducibility, and we note this as a limitation of our method.

We apply our full pipeline including segmentation and metrics extraction to each sweep in the clinical study set. Similarly, we apply Quader's [61] full pipeline to the same set of sweeps to extract α_{3D} and FHC_{3D} with their proposed algorithms that use CSPS and RFCs for segmentation. Additionally, we manually inspect and discard any sweeps according to the scan adequacy labeling procedure described in §5.1. If all the sweeps within an exam are discarded, we discard that exam completely, and subsequently do not include that hip in our final analysis. For each exam (containing 1-4 adequate sweeps), the assigned dysplasia metric is the average of that dysplasia metric across all the remaining adequate sweeps in that exam. Finally, each remaining exam is assigned the following five metrics:

- α_{3D} using our proposed methods
- α_{3D} using Quader's methods [61, 63]
- FHC_{3D} using our methods

Table 4.1: Results showing inter-exam, intra-sonographer ICC of our proposed pipeline for computing α_{3D}, FHC_{3D} and OCR vs. the state-of-the-art methods (n=42 hips).

	Method	**ICC (95% CI)**	**p-value (H1: ours > other method)**
$\boldsymbol{\alpha}_{3D}$	Quader's [61]	0.78 (0.62, 0.88)	0.03
	Ours	0.87 (0.77, 0.93)	-
$\mathbf{FHC}_{3D}$	Quader's [61]	0.68 (0.47, 0.81)	0.006
	Ours	0.84 (0.72, 0.91)	-
OCR	Quader's [61]	-	-
	Ours	0.74 (0.58, 0.86)	-

- FHC_{3D} using Quader's methods [61, 64]
- OCR using our methods

Given the inter-exam pairs for each of these 5 metrics, for each hip, we can compute the inter-exam reproduciblity using the ICC. Following the guideline for selecting ICC by Koo [38], we select a two-way mixed effects model, based on single measurement, with absolute agreement definition. In addition to ICC, we also report the test-retest Standard Deviation (SD)s for the different methods as another measure of reproducibility.

4.3.4 Results and Discussion

Starting from 483 sweeps (from 60 hips) in the clinical study set, 317 sweeps (from 42 hips) remain after we discard all inadequate sweeps. We report ICC for the DDH metrics with the different methods in table 4.1. Further, we report the the SDs in table D.1.

Comparing to the literature

To the best of our knowledge, there are only two other techniques in the literature that were proposed for extracting DDH metrics from 3D-US of the neonatal hip, and these are Quader's [61, 63, 64] and Zonoobi's [82]. We had access to Quader's

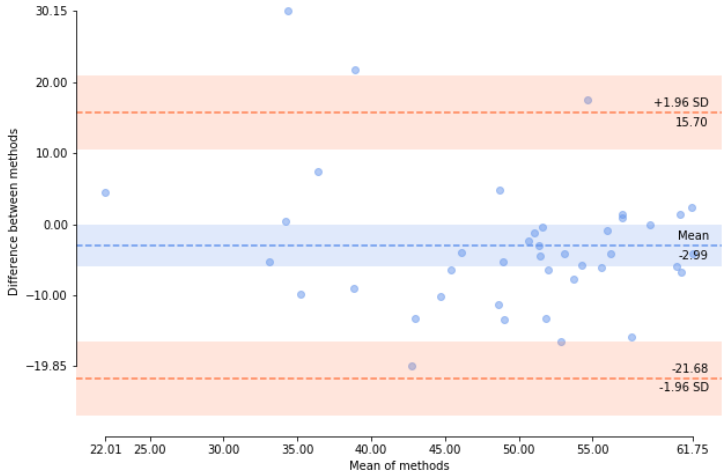
Table 4.2: Qualitative plausibility analysis of adequate sweeps in which a large discrepancy between our methods and Quader's (SOTA) was detected. Numbers reported are percentages of successful and plausible measurements out of 51 sweeps visually inspected. Inspection was performed by an unbiased rater other than the author of this thesis.

		% Plausible
α_{3D}	**Quader's [61, 63]**	76%
	Ours	100%
FHC_{3D}	**Quader's [61, 64]**	57%
	Ours	100%

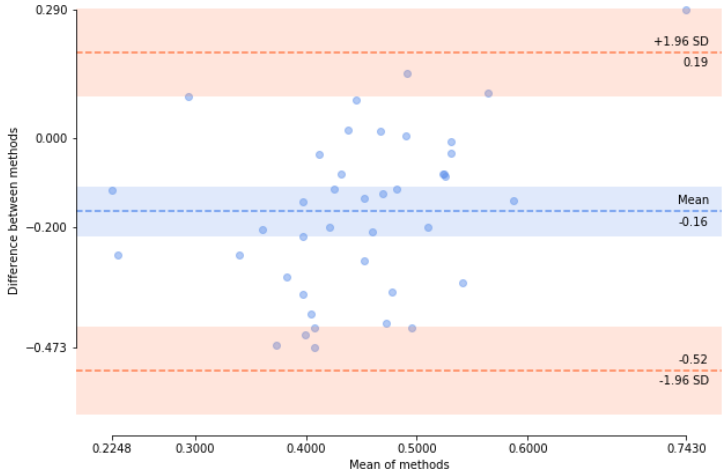
[61] code, so we were able to compare our methods to theirs directly on the same dataset. We use the R implementation of ICC in the IRR package [9], which performs a one-sided test against a specified null hypothesis, and these are the p-values we report in the ICC table 4.1. Our results show that across all three of the reported metrics, our proposed methods are more reproducible than Quader's [61]. We can only compare our methods to Zonoobi's [82] methods indirectly as we do not have access to their code, so we can only compare our results to the numbers reported in their publication. We first note the following differences in Zonoobi's methods: 1) they used a different ICC definition (two-way mixed effects, consistency, single rater; 3,1); 2) they averaged the measures from multiple sonographers at each exam; 3) instead of a single α_{3D} measure, they report α_{3D}-posterior and α_{3D}-anterior. With these differences, Zonoobi reports an inter-exam ICC on a set of 60 hips for α_{3D}-posterior, α_{3D}-anterior, and OCR of 68%, 62%, and 50%, respectively. Comparing to our ICC results for α_{3D} and OCR of 87% and 74%, respectively, it appears that our methods may be more reproducible, but this is yet to be confirmed in a direct comparison in future work.

Comparing the metrics to each other

Following Koo's guidelines [38] for ICC, our α_{3D}'s reproducibility is good to excellent, FHC_{3D} is moderate to excellent, and OCR is moderate to good. α_{3D} and FHC_{3D} are virtually tied, and they both significantly outperform OCR.



(a) α_{3D} $(^{\circ})$



(b) FHC_{3D} (unitless ratio)

Figure 4.16: Bland-altman plots showing large discrepancies between our metrics and Quader's metrics (n=42 hips).

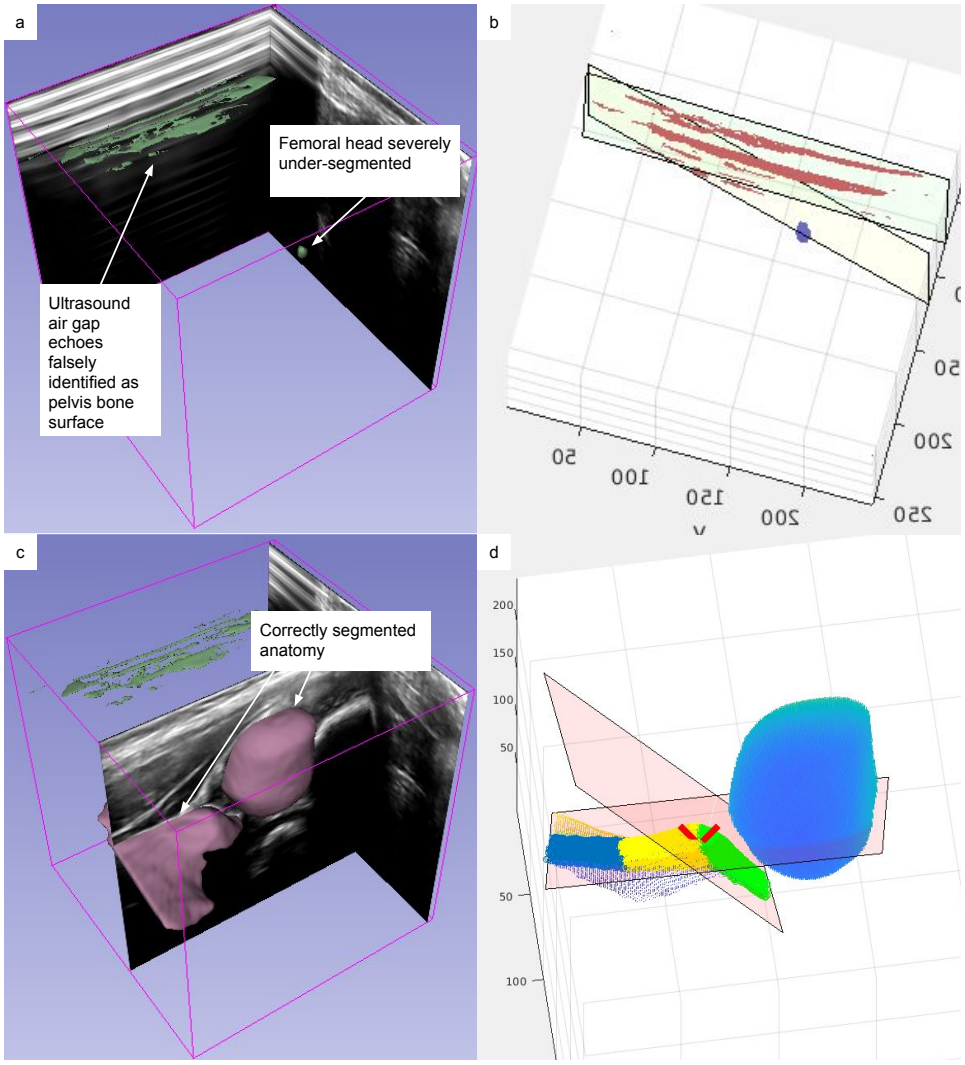


Figure 4.17: Example showing failure with Quader's SOTA CSPS-based method for pelvis bone surface segmentation and α_{3D} measurement [63]. a) Incorrectly segmented pelvis bone surface with Quader's method shown in green. b) The corresponding fitted planes resulting in an implausible α_{3D} measurement. c) The same sweep correctly segmented with 3D-U-Net shown in red. d) The corresponding fitted planes and plausible α_{3D} measurement.

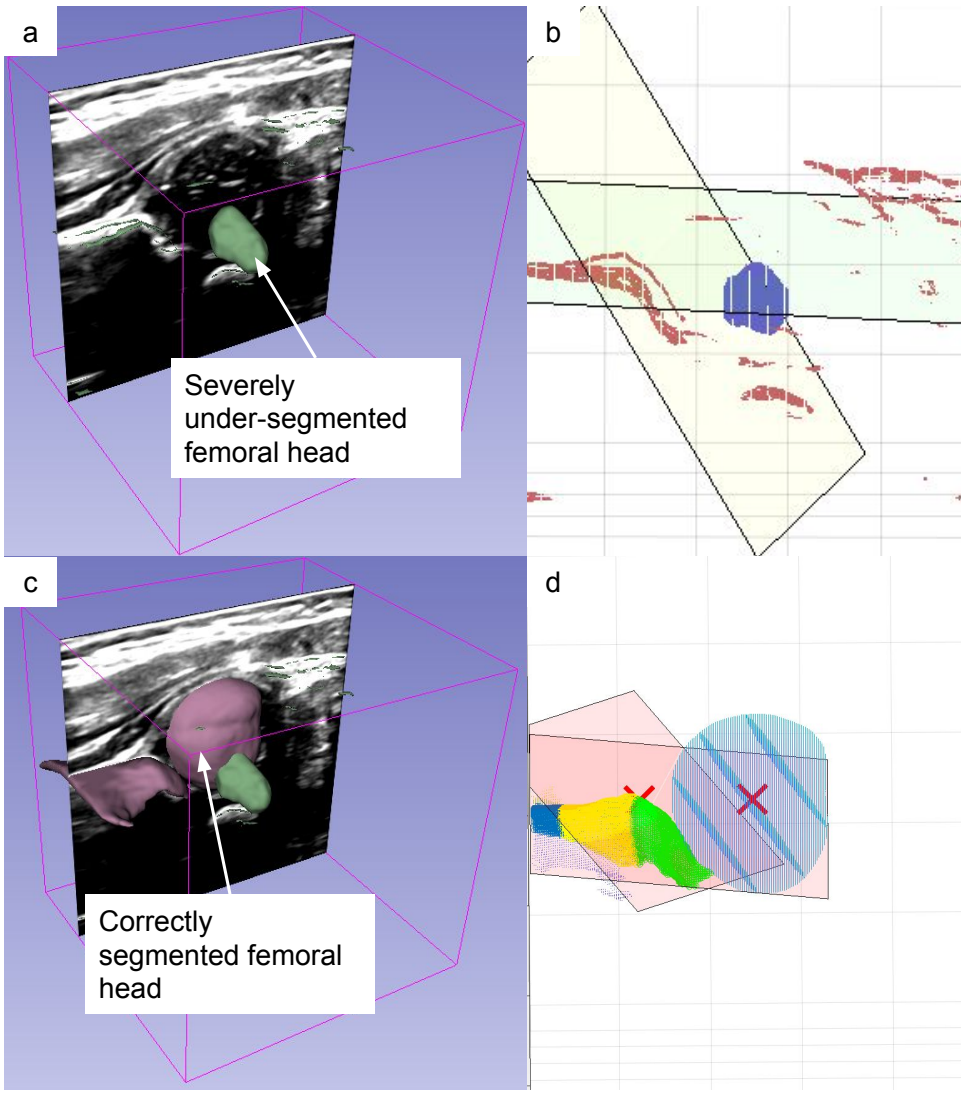


Figure 4.18: Example showing failure with Quader's SOTA RFC-based method [64]. a) Incorrectly segmented femoral head with Quader's method shown in green. b) The corresponding fitted planes resulting in an implausible FHC_{3D} measurement. c) The same sweep correctly segmented with 3D-U-Net shown in red. d) The corresponding fitted planes and plausible FHC_{3D} measurement.

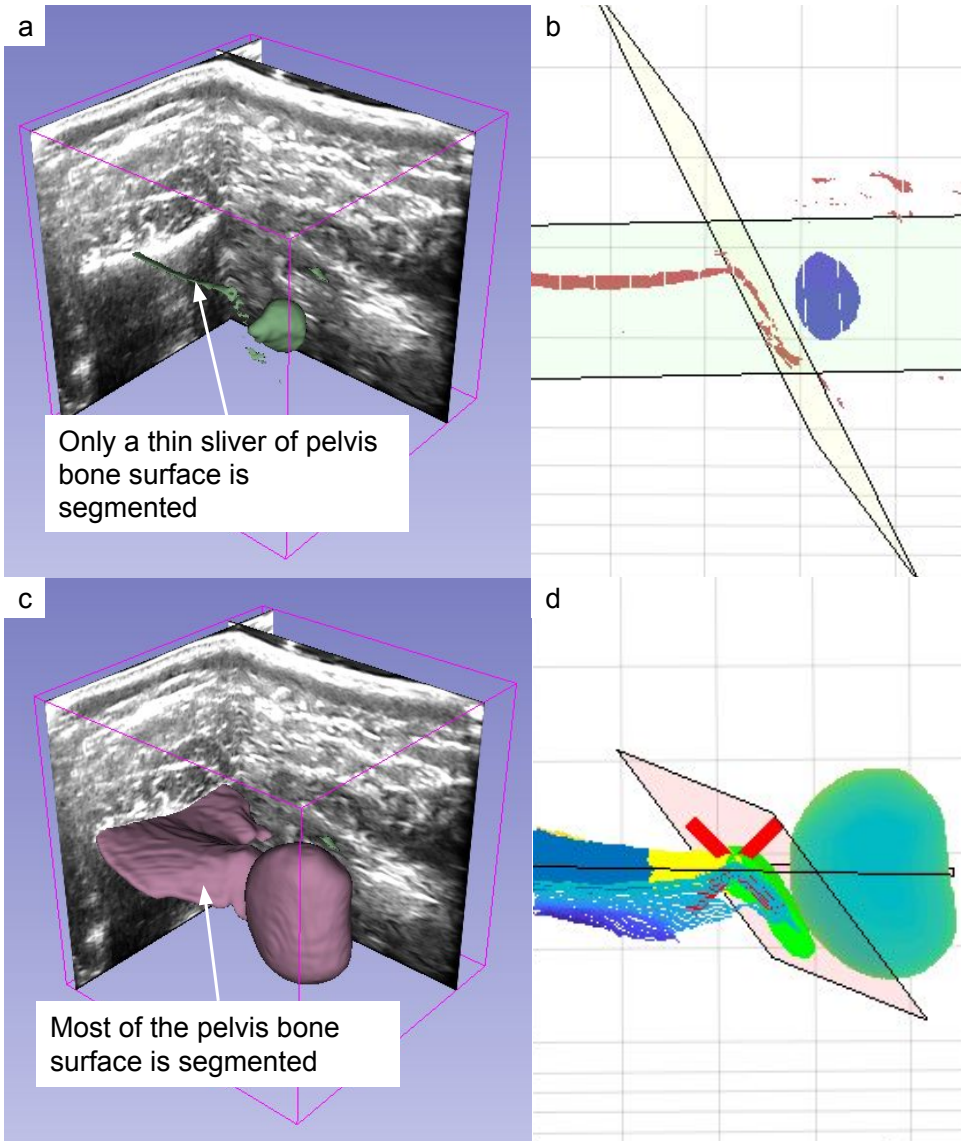


Figure 4.19: Example showing questionable case with Quader's SOTA CSPS-based method [63]. a) Quader's CSPS segmentation method only captures a very thin silver of the overall pelvis bone surface. b) The corresponding fitted planes resulting in a questionable α_{3D} measurement. c) The same sweep correctly segmented with 3D-U-Net shown in red. d) The corresponding fitted planes and plausible α_{3D} measurement.

Large Discrepancy Analysis and Failure Cases

On the set of 42 hips, we inspect cases where there are large discrepancies between our methods and Quader's. Since we do not have a trusted *gold standard* measurement of DDH in these participants due to the clinical standard 2D-US being unreliable, we cannot directly judge which of the two methods is closer to the true value, or in other words which method is more accurate. In these cases, we visualize the segmentation and plane-fitting outputs of Quader's methods and ours, to determine if we can make any conclusions which meathod is more plausible.

First, we discard inadequate sweeps as described in §5.1. Of the remaining adequate sweeps, we identify how many such large discrepancy cases exist with Bland-Altman plot analyses. For both α_{3D} and FHC_{3D}, we visualize the discrepancies on Bland-Altman plots as shown in figure 4.16. Note the relatively high 1.96SD ranges of -22° to 16° for α_{3D}, and -52% to 19% for FHC_{3D}, suggesting a high level of discrepancy between the two methods that warrants further investigation.

Based on these Bland-Altman plots, we select rough cut-offs of -10° to 10° for α_{3D}, and -30% to 10% for FHC_{3D}, and deem any cases with an absolute difference (i.e. difference between our metrics and Quader's metrics) beyond these thresholds as a *large discrepancy*. Based on this, we identified 51 adequate sweeps (from 7 hips) with large discrepancy.

An unbiased (i.e. not involved in any of the work done in this thesis or Quader's work, and does not have conflicts-of-interest) engineering student in our lab is asked to inspect each such case (adequate and large discrepancy). This individual is asked to judge the overall plausibility of the α_{3D} and FHC_{3D} measurements for the two contrasted methods based on the perceived quality of the segmentation and plane-fitting. We report the results of this analysis in table 4.2.

From these results, we can see that in all cases in the inspected dataset our proposed methods for α_{3D} and FHC_{3D} are always plausible, whereas measurements with Quader's methods are in many cases implausible. The most common reason for α_{3D} measurement implausibility with Quader's method was failure to segment the pelvis bone surface, instead falsely segmenting other regions (e.g. soft tissue or air gaps) and identifying them as the pelvis bone surface (7/51 such cases), and we

show an example of this in Figure 4.17. For FHC_{3D}, the rater observed many cases in which Quader's RFC severely undersegmented the femoral head, resulting in an implausible measurement (17/51 such cases), and we show an example in Figure 4.18. In addition to these clearly implausible cases, the rater reported observing some borderline (questionable) cases in which Quader's CSPS-based segmentation method resulted in only a very thin sliver of the total pelvis bone surface being segmented, for example see Figure 4.19. Finally, the rater reported 3/51 cases in which Quader's program failed and quit with an error before completing the segmentation and measurement, resulting in outputs of 0 for α_{3D} and FHC_{3D}.

4.3.5 Conclusions

We proposed a new algorithm for extracting DDH metrics including α_{3D}, FHC_{3D}, and OCR from segmented neonatal hip 3D-US. We showed that our methods produce higher inter-exam, intra-sonographer ICCs compared to the SOTA methods proposed by Quader [61, 63, 64]. It also appears that our methods may be more reproducible than the semi-automatic method proposed by Zonoobi [82], although this is yet to be determined in a direct comparison in future work. Further, we showed that in cases with large disagreement between our methods and Quader's, our methods appear to produce more plausible results and are more robust to failure.

Chapter 5

Automatic Adequacy Assessment with 3-Dimensional Convolutional Neural Networks

As described in Ch.1, Paserin's [56, 58] was the only work to explore automatic adequacy assessment for 3D neonatal hip ultrasound that made use of 3D data from adjacent slices. Paserin introduced this in the form of an RNN model that classifies whether or not a given volume is adequate based on the following criteria defined by Paserin in collaboration with a radiologist:

- The femoral head, a hypo-echoic spherical structure, should be fully present and seen growing and shrinking in size across the encompassing slices
- The ilium must appear as a straight, horizontal, hyper-echoic line
- The acetabulum must be present and appear continuous with the iliac bone
- Presence of ischium
- Presence of labrum
- All of these features should be collectively present within an adequate volume, but they do not necessarily all need to be present within any single slice

Paserin's RNN was able to emulate the radiologist's adequacy assessment quite well, achieving a reported AROC of 83% on a test set of 20 volumes. However, the completeness of the *criteria* themselves was not validated. Specifically, the effect of the choice of adequacy criteria on the reproducibility of α_{3D} and FHC_{3D} was not tested. In this chapter, we evaluate if using Paserin's proposed criteria can improve DDH measurement. Further, we propose our own adequacy criteria and compare these with the criteria proposed by Paserin, evaluating these against DDH metric reproducibility. Lastly, we show how 3D-CNNs can be used to automate adequacy classification with our proposed criteria.

5.1 Labeling with New Criteria

In §4.3.3, we described a clinical evaluation study in which a rater (the author of this thesis) inspected the full set of 483 sweeps from 34 participants, and judged which sweeps were adequate for DDH measurement. Having gained access to high quality segmentations of the hip anatomy with our trained models from the previous chapter, we could now see patterns in the anatomy which were not previously apparent, and consequently gained an improved understanding of the overall shape. Given this newfound understanding, we suggest that Paserin's adequacy criteria could be improved, and hypothesize that criteria that are more selective could ultimately improve the reliability of the DDH measurement. As such, we did not strictly adhere to Paserin's criteria in the labeling process. Instead, to label the 483 sweeps, the rater (author of this thesis) was asked to only answer the following simplified question for every sweep: "Is the sweep adequate for α_{3D} and FHC_{3D} measurement?". The rater was given a choice of answering the question as "yes", "maybe" (if not sure), or "no", after visualizing simultaneously 4 views: 1) B-mode coronal view cine, 2) B-mode sagittal view cine, 3) B-mode transverse view cine, and 4) the segmented anatomy as point clouds. Using this procedure, 317 sweeps were labelled as "yes", 101 as "maybe", and 65 as "no".

Retrospectively, the following are the reasons we most often observed for which we rejected a sweep:

- The ilium is fully or partially beyond the Field-of-View (FOV) of the probe, and this is usually caused by the probe being positioned too inferiorly. If

much of the ilium surface is missing, then we cannot be certain that the fitted plane represents the full ilium surface, and consequently is not adequate for α_{3D} or FHC_{3D} measurement. (e.g. see figure 5.1)

- Similarly, the femoral head is partially or fully beyond the FOV of the probe, and this can be caused by the probe being positioned too anteriorly or posteriorly. If the femoral head is occluded, we cannot make a FHC_{3D} measurement.

- There is clear movement artifact. This can be usually observed in the sagittal view and appears as a "smudge". It is also visible when playing a cine of the coronal view as we see the femoral head abruptly moving superiorly and inferiorly in the cine. (e.g. see figure 5.2)

- The ilium and acetabulum are present and planes can be fitted to them, but their segmented surface area appears smaller than other high-quality, "adequate" examples. This can be caused by the probe pose or rotation deviating significantly from the optimal position. (e.g. see figure 5.3)

We highlight the following differences compared to Paserin's criteria. With our criteria:

- We ignore the labrum and ischium, as these are not relevant for measuring α_{3D} and FHC_{3D}.

- We emphasize movement artifact, whereas this was not included in Paserin's criteria.

- The ilium is treated as a plane (which can be tilted), whereas Paserin's criteria specifies that it must be a horizontal line (and we assume that this is based on the standard plane as judged by the rater).

- Beyond the simple presence of certain anatomy viewed in the B-mode views, we emphasize probe positioning and image quality in terms of the *shape* and *surface* area of the segmentation, which can be more clearly observed in the 3D segmentation view, but is not obvious when looking at only the sagittal, coronal, and transverse B-mode cines.

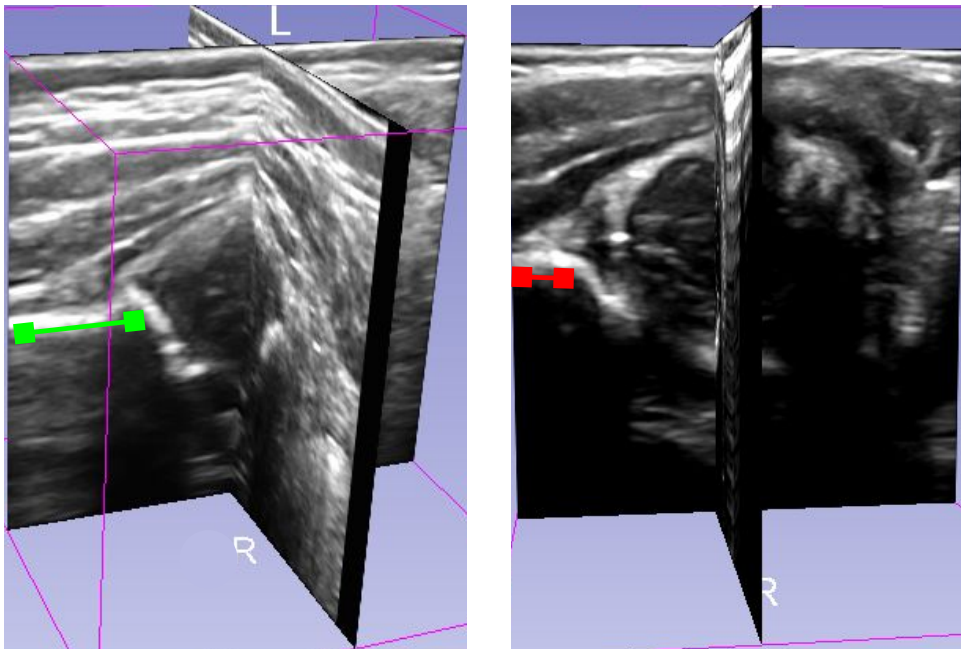


Figure 5.1: Example showing a sweep that was deemed "inadequate" because the ilium appears to be beyond the FOV of the scan due to the probe being positioned too inferiorly from the optimal position (right), and an "adequate" volume with ilium fully within the FOV for comparison (left).

5.1.1 Evaluation Scheme

To evaluate and compare our new criteria with Paserin's, we apply Paserin's RNN model [56, 58] to the same clinical evaluation dataset described in §4.3.3, and discard sweeps labelled as inadequate by this model. We again compute the following metrics for all 483 sweeps in the set:

- α_{3D} using our methods described in Ch.4
- FHC_{3D} using our methods described in Ch.4
- OCR using our methods described in Ch.4

And we compare the inter-exam, intra-sonographer ICC of the following three sets:

- No-Discard: The full clinical evaluation set without discarding any sweeps

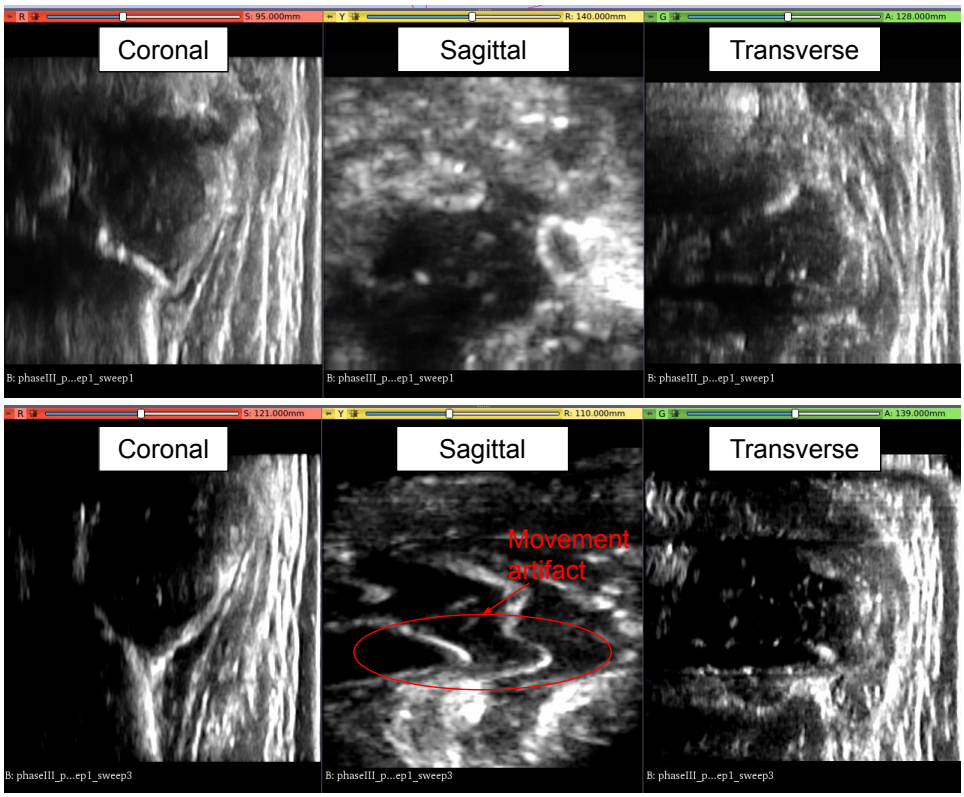


Figure 5.2: Example of a sweep deemed "inadequate" because of movement artifact that can be seen as a "smudge" in the sagittal view (lower row), and for comparison we show the sagittal view of an "adequate" volume (top row).

- Paserin-Criteria: A distilled set, discarding "inadequate" sweeps based on the predictions from Paserin's RNN model [56, 58], and we consider the predictions of this model to be an approximation of Paserin's adequacy criteria
- Our-Criteria: A distilled set, discarding "inadequate" sweeps based on the new criteria describe in the previous §5.1

5.1.2 Results and Discussion

We report the results in table 5.1. First, we note that only 17 sweeps out of 483 were labelled as inadequate and rejected with the RNN, compared to 166 labeled

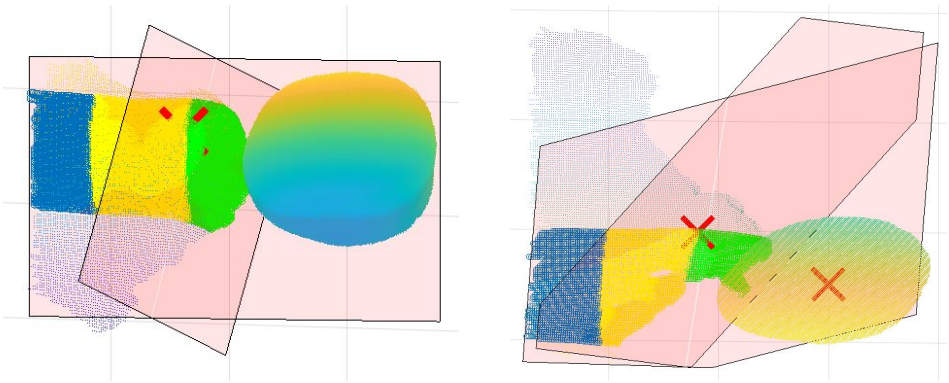


Figure 5.3: Example of a sweep deemed borderline adequate ("maybe") on the right, due to the labeler's perception that the probe was not positioned optimally. Note the shape and reduced area of the ilium (green) and acetabulum (yellow) surfaces used for α_{3D} in the sweep on the right, compared with the high-quality sweep on the left. This is potentially due to the probe being slightly tilted (roll around x-axis) or translated (along z-axis) away from the optimal position.

Table 5.1: Comparing inter-exam, intra-rater test-retest ICC with different adequacy criteria. The number of sweeps remaining after discarding inadequate sweeps n is shown in parentheses beside each column header. The 95% CI is reported in parentheses next two each ICC number.

	No-Discard (n=483)	**Paserin-Criteria (n=466)**	**Our-Criteria (n=317)**
$\boldsymbol{\alpha}_{3D}$	0.65 (0.48,0.78)	0.67 (0.50,0.79)	0.87 (0.77,0.93)
FHC$_{3D}$	0.74 (0.61,0.84)	0.75 (0.62,0.85)	0.84 (0.72,0.91)
OCR	0.67 (0.50,0.79)	0.68 (0.52,0.80)	0.74 (0.58,0.86)

as inadequate (in the "maybe" or "no" category) and rejected with the proposed criteria. Further, the ICCs across all three metrics appear to be higher in the Our-Criteria set compared to the other two, whereas the Paserin-Criteria set appears to be tied with the No-Discard set. This suggests that our adequacy criteria are more selective compared to Paserin's RNN, and that this selectivity appears to increase test-retest reproducibility.

Table 5.2: Adequacy train and test sets class distribution.

	Yes=1	**Maybe=0.5**	**No=0**	**Total**
Train Set	100	91	145	336
Test Set	81	30	25	136
Total	181	121	170	472

5.2 Automatic Adequacy Classification with 3D-CNNs

As was described in Ch.1, the overall objective of this thesis is to develop a fully automatic system that is user-independent, so in this section we attempt to automate the adequacy classification step that was performed manually as was described in §5.1.

5.2.1 Classification Model

Inspired by the high accuracy and speed of CNNs proposed by Paserin [56–58] for this task, we also propose to use CNNs for this task. However, similar to our methods proposed for femoral head sphere regression presented in §4.2, we propose to use 3D-CNNs that can fully capture the 3D information in the full volume, compared to the 2D-CNNs that can only use limited information from *single frames* and RNNs that can only use information from a *few adjacent slices*. Again, we experiment with 3D-ResNet-50 and 3D-DenseNet-121 based on promising performance with on a video classification task reported in the literature [18–20], and given our time and hardware limitations.

5.2.2 Labeling Data for CNN Training

As depicted in figure 2.3, for training and testing our models, we use an expanded set of volumes from all 118 participants in our full dataset. We assign 336 sweeps from 84 participants to the training set and 136 sweeps from 34 participants to the test set, totaling 472 sweeps. The same rater (author of this thesis) labelled this set of training and test data using the procedure described in §5.1. The class distribution for the train and test sets is illustrated in figure 2.3 and is summarized in table 5.2.

5.2.3 Training

Based on observations from preliminary cross-validation experiments, we choose 3D-DenseNet-121 over 3D-ResNet-50 for our classification task. We train three models based on 3D-DenseNet-121 with the following differences:

- Model 1 [B_Y/N]
  - Input: single channel B-Mode input only
  - Output: binary class label, 1 for adequate and 0 for inadequate
  - Training set class distribution: All sweeps labelled as "yes" are assigned to one class (adequate), all sweeps labelled as "no" are assigned to the other class (inadequate), and all sweeps labelled as "maybe" are ignored and not included in the training.
- Model 2 [B+Seg_Y/N]
  - Input: 3-channel input with B-mode in one channel, 3D-U-Net binary mask prediction of the pelvis bone surface in the second channel, and 3D-U-Net binary mask prediction of the femoral head in the third channel.
  - Output: binary class label, 1 for adequate and 0 for inadequate
  - Training set class distribution: All sweeps labelled as "yes" are assigned to one class (adequate), all sweeps labelled as "no" are assigned to the other class (inadequate), and all sweeps labelled as "maybe" are ignored and not included in the training.
- Model 3 [B+Seg_Y/M/N]:
  - Input: 3-channel input with B-mode in one channel, 3D-U-Net binary mask prediction of the pelvis bone surface in the second channel, and 3D-U-Net binary mask prediction of the femoral head in the third channel
  - Output: binary class label, 1 for adequate and 0 for inadequate

- Training set class distribution: All sweeps labelled as "yes" are assigned to one class (adequate), all sweeps labelled as "maybe" or "no" are assigned to the other class (inadequate).

For all three models we select the same training hyperparameters:

- Batch size: 1
- Learning rate: 0.0001
- Optimizer: Adam [37]
- Loss: BCE
- Epochs: 150
- Augmentation as described in §4.2.2
- Input size: $100 \times 100 \times 100$ pixels
- Regularization: dropout with rate 50%

5.2.4 Testing

We compare the performance of the three models using the AROC on the test set. Although not directly comparable as it was trained on a different training set, for completeness we also report the Paserin's RNN AROC against the same test set. Since our models output is binary (adequate/inadequate), whereas the test set includes three classes (yes/maybe/now), we compute the AROCs on two subsets of the test set:

- Test subset 1: ignoring the "maybe" test cases (this reduces the test set from 136 to 111 sweeps), and
- Test subset 2: assigning all "maybe" cases to the "inadequate" class

Further, we repeat the inter-exam, intra-sonographer ICC analysis on the clinical evaluation sweep (483 sweeps), but this time using the predictions from the three contrasted models to discard inadequate sweeps. We report ICCs for the

Table 5.3: AROC scores of three contrasted models when applied on the test set. In the first row we ignore sweeps in the test set labelled as "maybe". In the second row we assign all sweeps labelled as "maybe" to the "inadequate" class.

Tested on:	**RNN**[58]	**B_Y/N**	**B+Seg_Y/N**	**B+Seg_Y/M/N**
Subset1 (Y=1, M=ignore, N=0)	0.49	0.89	0.9	0.84
Subset2 (Y=1, M=0, N=0)	0.49	0.75	0.84	0.83

Table 5.4: Inter-exam, intra-sonographer ICC with the proposed CNNs. n is the number of remaining sweeps after "inadequate" sweeps are discarded. 95% CIs are shown in parentheses.

	B_Y/N (n=335)	**B+Seg_Y/N (n=391)**	**B+Seg_Y/M/N (n=345)**
$\boldsymbol{\alpha}_{3D}$	0.79 (0.65,0.88)	0.76 (0.62,0.86)	0.73 (0.50,0.86)
$\mathbf{FHC}_{3D}$	0.82 (0.70,0.90)	0.77 (0.64,0.87)	0.81 (0.64,0.91)
OCR	0.60 (0.40,0.76)	0.72 (0.56,0.83)	0.67 (0.41,0.83)

DDH metrics given that inadequate sweeps are discarded with the three trained 3D-DenseNet-121 models, and compare these with the previously proposed strategies: No-Discard, Paserin-Criteria (RNN), and Our-Criteria (manually labeled) summarized in table 5.1.

5.2.5 Results and Discussion

The AROC scores in table 5.3 show that all three 3D-DenseNet-121 models have mostly learned to emulate the labeler's ability to predict adequacy based on our proposed criteria. Paserin's RNN, in comparison, has a very poor score of 49%, indicating that it is not good at predicting adequacy based on our new criteria, although Paserin reported a high AROC of 83% on a test set of 20 volumes labeled with their own criteria. Model1 [B_Y/N] performs well (AROC of 89%) when tested on test subset1 that does not include "maybes", but performs relatively poorly (AROC of

75%) when tested on subset2 that includes "maybes". This suggests that this model is less selective and might classify borderline "maybe" cases as "adequate", which presents a risk in a real-world scenario where it is safer to be more selective and repeat acquisition for borderline cases, as repeated acquisitions are low cost and fast. In contrast, Model2 (B+Seg_Y/N) and Model3 (B+Seg_Y/M/N) which additionally take as input the segmentation binary masks, scored higher AROCs of 84% and 83% on Subset2. This suggests that segmentation data may provide additional useful information that can help determine adequacy in uncertain cases. Models 2 and 3 appear to be more selective compared to Model 1, suggesting that they are potentially safer for use in real-world scenarios.

Considering the ICC scores in table 5.4 and comparing these to scores in table 5.1, in general we see worse test-retest reproducibility with the 3D-DenseNet-121 compared to the human-labelled "Our-Criteria" set. This is probably explained by over-fitting to the training data and a relative decrease in accuracy on unseen test data, despite using augmentation and dropout regularization. This could be improved with more data which we did not have. Further, based on the reported ICC scores, we do not see a clear pattern on which of the three 3D-DenseNet-121 models performs the best across all three DDH metrics. However, in general we see that models 2 and 3 appear to score as good or better than the No-Discard and Paserin-Criteria sets, suggesting their potential to assist in improving DDH measurement in a real-world scenario.

5.2.6 Conclusions

We proposed new adequacy criteria and compared these criteria to Paserin's [56, 58], the closest work to ours, and to the best of our knowledge the only other work on scan adequacy for 3D-US for DDH. We showed that our newly proposed criteria capture more relevant information and are more selective. Due to this selectivity, we showed that using these criteria to discard inadequate sweeps improves reproducibility of α_{3D} and FHC_{3D} measurement, suggesting that using the proposed adequacy criteria may reduce misdiagnosis (to be confirmed in future work). Further, we evaluated 3D-DenseNet-121 to automate adequacy classification with our criteria. We show that our trained models capture more information and are more

selective compared to the RNN, but we did not show conclusively which of the three proposed training regimes for 3D-DenseNet-121 is the best.

Limitations

We can only make limited conclusions due to the limitations of this study. These include:

- The adequacy criteria we proposed and used, although they showed improvements in ICC, are not precisely defined and remain subjective.
- There is a data imbalance in our train and test sets. For example 43% of the training data was labelled as "no", whereas only 18% of the test data was labelled as "no". It appears that our sonographers became better at acquiring adequate images as time progressed. This can also be seen in the last row of figure 2.3, of which the x-axis is ordered chronologically, we see that as time progresses we get more "yes" labels, and fewer "maybe" and "no" labels.

Future Work

Future work will focus on improving the adequacy criteria, the models, and experimental methodologies. Suggestions for future work include:

- More precisely and quantitatively defining adequacy criteria to improve labeling reproduciblity.
- Having more than one labeler, which allows testing reproducibility of the criteria and reduces labeler bias.
- Using novice sonographers, ideally having a new sonographer for every scan, which would reduce the learning effect described in the limitations.
- Mitigating the class imbalance in the training process, by reorganizing the train and test sets, re-sampling, or re-weighting training examples for example with Focal Loss [41].
- In our analysis we simply discarded "inadequate" cases. In a real clinical scenario, this presents a real problem, as an inadequate case means that the

DDH diagnosis must be made based on information from different tests, potentially forcing the clinician to revert to standard techniques including 2D-US and clinical examination. As these techniques are unreliable, as we discussed in Ch.1, this puts the patient at risk of misdiagnosis. Future work should incorporate point-of-care adequacy feedback, and assess how many of all the "inadequate" cases can be adequately re-acquired.

- An interesting avenue that could be explored in future work is incorporating a built-in uncertainty measure into the network that can tell the user if the network is uncertain of its prediction. This additional uncertainty information may potentially help identify the "maybe" cases.

Chapter 6

Discussion and Conclusions

6.1 Revisiting Research Questions and comparing the State-of-the-Art

Here we reiterate the research questions presented in §1.4, and outline conclusions made for each RQ based on work presented in the previous chapters, in the process comparing our proposed methods to the SOTA previously presented in the literature.

6.1.1 Research Question 1

> *Can* CNN*s be trained to segment the pelvis bone surfaces, including the ilium and acetabulum, in neonatal hip* 3D-US*? Would the predictions produced by such* CNN*s more closely resemble human labels, as compared to existing* SOTA *methods such as* CSPS*?*

In Ch.3, we trained U-Net [67] on 439 2D slices for segmenting the pelvis bone surface in neonatal hip 3D-US. We chose U-Net because of its proven success for medical image segmentation [30] and ability to learn from very few images. In our comparison, we mainly compare with CSPS as it is the only fully automatic method that was applied to 3D-US volumes for DDH, to the best of our knowledge, and we consider to be the SOTA for our application. When tested on 103 previously unseen 2D slices, U-Net achieved a DSC of 86%, outperforming CSPS+ROI [61, 62] which achieved a DSC of 81%.

In §4.1, we further experimented with 3D-U-Net [6], which uses 3D convolution kernels as opposed to U-Net which uses 2D convolutions, hypothesizing that incorporating 3D information may improve the segmentation accuracy. We trained 3D-U-Net on a set of 64 volumes. When tested on 52 volumes, we report DSCs of 85% an 91% for 3D-U-Net and U-Net, respectively, but very different CDI scores of 76% and 24% with 3D-U-Net and U-Net, respectively. These results, supported by our visual observations, suggest that 3D-U-Net is less likely to produce false positive segmentations (islands) that are distant from the bone surface, but that both methods can capture most of the bone surface. Importantly, these are much higher than the scores achieved by the SOTA CSPS method, which achieved DSC of 26% and CDI of near 0%.

We conclude that CNNs can be trained to segment the pelvis bone surfaces in neonatal hip 3D-US, and that the resulting segmentations more closely resemble human labels as compared to the SOTA.

6.1.2 Research Question 2

> *Can* CNN*s be trained to locate the femoral head in neonatal hip* 3D-US*? Would the predictions produced by such* CNN*s more closely resemble human labels, as compared to existing* SOTA *methods such as Quader's* RFC *[61, 64]?*

In §4.2, we trained 3D-U-Net [6] to segment the femoral head in neonatal hip US. We compare 3D-U-Net to an RFC-based method introduced by Quader [61, 64], which is to the best of our knowledge the only method previously presented in the literature for fully automatic segmentation of the femoral head in 3D-US of the neonatal hip, and which we consider the SOTA for this application. When tested on a set of unseen 53 volumes, 3D-U-Net localized the femoral head with CED 1.42 mm and RAE 0.46 mm. This more closely matched the human label compared to the RFC [61, 64], which achieved 3.90 mm CED and 2.01 mm RAE on the same test set. We conclude that CNNs can be trained to locate the femoral head in neonatal hip 3D-US, and that the predictions produced by these CNNs more closely resembles the human labels as compared to the SOTA.

6.1.3 Research Question 3

Can we develop automatic methods for extracting α_{3D} and FHC_{3D} metrics with our improved segmentations that are at least as reproducible as the previously proposed methods [61, 63, 64]? Can we show that our proposed methods are at least as robust and plausible as these previously proposed methods?

In §4.3, we presented algorithms for automatically extracting dysplasia metrics including α_{3D}, FHC_{3D}, and OCR from the segmented neonatal hip 3D-US volumes. Comparing to the literature, to the best of our knowledge, there are only two other systems which use 3D-US for DDH diagnosis that are comparable to our system, and both were developed simultaneously. One system, which was developed by Zonoobi et al. [82], is semi-automatic, requiring seed-point inputs from the user, and measures the α_{3D}-anterior, α_{3D}-poster, and OCR metrics. The other system, developed in our lab by Quader et al. [61, 63, 64], is fully automatic, uses CSPS for bone surface segmentation, an RFC classifier for femoral head segmentation, and measures α_{3D} and FHC_{3D}. Of these two, Quader's method, being fully automatic, is closest to ours, so we consider this to be the SOTA for our application, so we compare directly to this method in our experiments.

On a clinical set of 42 hips, our method achieves inter-exam, intra-sonographer ICCs of 87%, 84%, and 74% for α_{3D}, FHC_{3D}, and OCR, respectively. On the same set of 42 hips, Quader's methods [61, 63, 64] achieved lower ICCs of 78% and 68% for α_{3D} and FHC_{3D}, respectively. Further, qualitative observations by an independent observer suggest higher plausibility, fewer failures, and improved robustness with our methods.

Based on our experiments in §4.3, we conclude that our methods are more reproducible (in the inter-exam, intra-sonographer setting), more robust, and more plausible than Quader's methods [61, 63, 64], the current SOTA for fully automatic measurement of DDH from neonatal hip 3D-US volumes.

6.1.4 Research Question 4

Are the current adequacy criteria proposed by Paserin [56] sufficient? Can we improve the criteria? Can we train new models for automating

classification based on the newly defined criteria?

To the best of our knowledge, Paserin's work [56–58] is the only other work in the literature that addressed the problem of adequacy classification of 3D-US, so we consider it the SOTA for this application. In §5, we propose a new set of *adequacy criteria* based on recent observations of segmented volumes, and compare these to Paserin's. When applied to a set of 483 volumes, only 317/483 cases were deemed adequate with the new criteria, whereas 466/483 were deemed adequate with Paserin's criteria (as approximated by the RNN [58]). Based on this, we conclude that our new criteria are more selective. Further, we report higher inter-exam, intra-sonographer ICCs when inadequate volumes are discarded with our criteria, for example 87% for α_{3D} vs. 67% when using the RNN for adequacy classification. These results, in addition to qualitative observations, suggest that the newly proposed criteria are more selective, and that this selectivity results in improved test-retest reproducibility of DDH measurement.

Further, we experimented with 3D-DenseNet-121 [18–20], for automating the adequacy classification based on the new adequacy criteria. Tested on an unseen test set of 136 sweeps, which was labelled based on the new criteria, 3D-DenseNet-121 achieved classification AROC of 84%, much higher than the RNN which achieved an AROC of 49%. With this improved selectivity, using 3D-DenseNet-121 for identifying and discarding inadequate sweeps, we observed higher ICCs compared to using the RNN for identifying discarding inadequate sweeps, but still lower than manual labeling (e.g. for α_{3D}, ICCs of 65% without discarding inadequate sweeps, 67% with the RNN, 76% with 3D-DenseNet-121, and 87% with manual inadequate sweep identification). We conclude that 3D CNNs show some promise towards this task, but can likely be much improved in future work.

6.2 Limitations

Overall, the biggest limitations of the work presented in this thesis include:

- Homogeneity and limited diversity in the data. For example, the data collected included only scans with the Ultrasonix 4DL14-5 probe, and from a sample of participants only from British Columbia. A known problem with

deep neural networks is their tendency to overfit to the training data, which is best mitigated by training with a diverse dataset, so our models have likely overfitted to our limited dataset. Their performance will likely deteriorate when used on data from other domains, such as different US probes, but this is yet to be determined with an expanded, more diverse dataset.

- Due to the lack of a reliable and trusted gold standard diagnostic technique for measuring hip socket depth, we do not report *accuracy* or *validity* of our methods, only test-retest *reproducibility*. Therefore, we cannot make any strong conclusions about our methods' validity beyond our qualitative observations that show plausibility of our proposed methods.

- The ICCs reported are inter-exam, *intra*-sonographer, and only two *expert* sonographers participated in the study. Therefore, conclusions about reproducibility cannot be generalized to *inter*-sonographer and *novice* user scenarios.

6.3 Future Work

Ultimately, the goal of this project is to develop an accurate, safe, and robust solution for DDH diagnosis. This device should additionally be optimized for cost, computational efficiency, and usability to facilitate wide-spread clinical adoption to reach as many participants as possible, and to reduce misdiagnosis rates globally. Building on work presented in this thesis, and considering these overarching goals, I recommend that future research should prioritize addressing the aforementioned limitations, as well as exploring new research avenues that would target these goals.

6.3.1 Domain Shift and Adaptation

Data from different domains: To address the problem of homogeneous data and domain shift, the first challenge would be to obtain more diverse data, for example from different probes, settings, and geographical regions. This is potentially possible with the help of clinical researchers at the International Hip Dysplasia Institute. This data would be crucial not only for evaluating, but also for improving

the accuracy of our models under domain shift. Here, different data scenarios may arise, which would require different solutions. For example, in the best-case scenario, we may get many images and many labels from different domains; or worse, we may get many images but few or no labels; or in the worst-case scenario, we may get few images and few labels. All of these are realistic scenarios, and present interesting research avenues.

Solutions for domain shift: Ultimately, the solution will depend on the available images and labels. In the scenario where many unlabelled images are provided from a new domain, unsupervised *domain adaptation* techniques can be used [75, 77]. A related approach is *neural style transfer*, whose main application has been artistic style transfer [34], but one can imagine each probe and setting combination as a different artistic style and apply the same techniques. In the scenario with many images but few or weak labels, *weakly-supervised* techniques, which use cheaper labels such as image-level tags to train segmentation networks, or *semi-supervised* techniques, which leverage a small number of strongly labelled images and many weakly-labelled images to cheaply improve segmentation, have been proposed [55]. In the more challenging scenario where images are scarce, new approaches such as *few-shot* learning [15], which aims to learn with very few labelled images, may be useful. Notably, few-shot techniques for segmentation are seemingly relatively under-studied compared to classification. The final solution will likely be a combination of such techniques as dictated by the available data.

Evaluation: Robustness to domain shift of proposed models and solutions could then be evaluated on a more diverse and heterogeneous dataset designed to test performance under domain shift. For example, solutions could be trained on data from one domain (e.g. with the Ultrasonix probe), and then tested on data from an unseen domain (e.g. different probe and settings). Depending on the task (e.g. segmentation or adequacy classification), performance could be measured and contrasted quantitatively with relevant metrics for the task (e.g. classification accuracy, Dice Score, etc.) on examples from the unseen domain.

6.3.2 Improved Clinical Study

To address the aforementioned limitations of *validity* (as opposed to reliability) and *intra*-sonographer ICC, an improved clinical study will likely be necessary in the future. Ideally, one could conduct a randomized clinical trial which would randomize patients into control and experimental groups, treat based on diagnosis with our 3D-US-based methods, and track clinical outcomes in the long run to assess sensitivity, specificity, and AROC of our proposed diagnostic techniques. However, this is likely not possible due to ethical considerations and resource limitations. Alternatively, one could make smaller modifications to future clinical studies to address the limitations, at least partially. For example, to assess *inter*-sonographer (as opposed to intra-sonographer) reproducibility, one could require each hip to be scanned by more than one sonographer. To address the question of *validity*, one could perhaps require patients to be scanned with a different imaging modality (e.g. MRI).

6.3.3 Detectability of Failure: Deep Learning with Uncertainty

Another interesting topic to be explored in future work is *uncertainty*. Patient safety is paramount in medical applications. Safety is a function of severity, probability, and detectability of failure. The biggest safety risk with using our AI-enabled US device is perhaps the risk of misdiagnosis. Work in this thesis has focused on reducing the *probability* of misdiagonsis. Further implementing a measure of uncertainty in our models would improve *detectability* of failure and potential misdiagnosis. This is perhaps especially important under scenarios of domain shift (e.g. different probe), in which an indication of low confidence can alert the operator that the model output cannot be trusted and that manual intervention may be necessary.

Uncertainty is an active area of research that has gained much attention recently. Many methods have been proposed to estimate uncertainty, perhaps most popular of which are Monte Carlo Dropout [8] and Bayes by Backprop [3]. A recent study [52] that compared the aforementioned techniques and others under domain shift concluded that quality of uncertainty degrades with increasing dataset shift regardless of method. It would be interesting to evaluate different uncertainty

techniques (e.g. Monte Carlo Dropout, BNNs, etc.) using the metrics proposed by Ovadia [52] on the DDH dataset, including Reliability Diagrams, Expected Calibration Error, and Entropy.

6.4 Clinical Impact and Significance

We have proposed a system for DDH diagnosis with 3D-US. CNNs played a key role in improving the segmentation and classification components of this system, ultimately improving the reliability, robustness, and usability of the system as a whole, as we showed in a limited clinical study. We hypothesize that these improvements will serve to improve the accuracy of DDH diagnosis in the clinic, reducing misdiagnosis rates, and consequently improving patient outcomes and reducing costs. Improved automation and usability of our system further serves to make this solution more attractive to clinicians, especially in low-resource settings that lack expertise, ultimately encouraging clinical translation of our system and consequently reaching more patients globally.

Bibliography

[1] 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*, 14(5): 775–783, may 2019. ISSN 1861-6410. doi:10.1007/s11548-019-01934-0. URL https://doi.org/10.1007/s11548-019-01934-0http://link.springer.com/10.1007/s11548-019-01934-0. → pages 12, 28, 34

[2] T. G. Barlow. EARLY DIAGNOSIS AND TREATMENT OF CONGENITAL DISLOCATION OF THE HIP. *The Journal of Bone and Joint Surgery. British volume*, 44-B(2):292–301, may 1962. ISSN 0301-620X. doi:10.1302/0301-620X.44B2.292. URL http://online.boneandjoint.org.uk/doi/10.1302/0301-620X.44B2.292. → page 6

[3] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra. Weight uncertainty in neural networks. *32nd International Conference on Machine Learning, ICML 2015*, 2:1613–1622, 2015. → page 89

[4] R. J. Cabin and R. J. Mitchell. To Bonferroni or Not to Bonferroni : When and How Are the Questions of America Society Bulletin Ecological. *America*, 81(3):246–248, 2010. → pages 43, 54

[5] A. Canziani, A. Paszke, and E. Culurciello. An analysis of deep neural network models for practical applications. *CoRR*, abs/1605.07678, 2016. URL http://arxiv.org/abs/1605.07678. → page 49

[6] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. In S. Ourselin, L. Joskowicz, M. R. Sabuncu, G. Unal, and W. Wells, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, pages 424–432, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46723-8. → pages 17, 36, 38, 46, 84

[7] C. Dezateux and K. Rosendahl. Developmental dysplasia of the hip. *The Lancet*, 369(9572):1541–1552, 2007. ISSN 0140-6736. doi:https://doi.org/10.1016/S0140-6736(07)60710-7. URL http://www.sciencedirect.com/science/article/pii/S0140673607607107. → page 1

[8] Y. Gal and Z. Ghahramani. Dropout as a bayesian approximation: Insights and applications. In *Deep Learning Workshop, ICML*, volume 1, page 2, 2015. → page 89

[9] M. Gamer, J. Lemon, I. Fellows, and P. Singh. Package 'irr', 2019. URL https://cran.r-project.org/web/packages/irr/irr.pdf. → page 63

[10] R. Ganz, M. Leunig, K. Leunig-Ganz, and W. H. Harris. The etiology of osteoarthritis of the hip: An integrated mechanical concept. *Clinical Orthopaedics and Related Research*, 466(2):264–272, 2008. ISSN 15281132. doi:10.1007/s11999-007-0060-z. → page 1

[11] D. Golan, Y. Donner, C. Mansi, J. Jaremko, and M. Ramachandran. Fully Automating Graf's Method for DDH Diagnosis Using Deep Convolutional Neural Networks. In *LABELS 2016*, volume 10008, pages 130–141, 2016. ISBN 978-3-319-46975-1. doi:10.1007/978-3-319-46976-8_14. URL http://link.springer.com/10.1007/978-3-319-46976-8http://link.springer.com/10.1007/978-3-319-46976-8{_}14. → page 13

[12] R. Graf. Fundamentals of sonographic diagnosis of infa[1] R. Graf, "Fundamentals of sonographic diagnosis of infant hip dysplasia," J. Pediatr. Orthop., vol. 4, no. 6, pp. 735–740, 1984.nt hip dysplasia. *Journal of Pediatric Orthopedics*, 4(6):735–740, 1984. ISSN 0271-6798. doi:10.1097/01241398-198411000-00015. → pages 2, 8, 9, 57

[13] R. Graf, S. Scott, K. Lercher, F. Baumgartner, and A. Benaroya. *Hip Sonography*. Springer Berlin Heidelberg, 2006. ISBN 978-3-540-30957-4. doi:10.1007/3-540-30958-6. URL http://books.google.com/books?id=x6Z8Sh8fJUEC{&}pgis=1http://link.springer.com/10.1007/3-540-30958-6. → pages xv, 3, 8

[14] R. Graf, M. Mohajer, and F. Plattner. Hip sonography update. quality-management, catastrophes - tips and tricks. *Medical Ultrasonography*, 15(4):299–303, 2013. ISSN 2066-8643. doi:10.11152/mu.2013.2066.154.rg2. URL https://www.medultrason.ro/medultrason/index.php/medultrason/article/view/772. → pages 8, 24

[15] A. Guha Roy, S. Siddiqui, S. Pölsterl, N. Navab, and C. Wachinger. 'Squeeze & excite' guided few-shot segmentation of volumetric images. *Medical Image Analysis*, 59, 2020. → page 88

[16] V. Gulati. Developmental dysplasia of the hip in the newborn: A systematic review. *World Journal of Orthopedics*, 4(2):32, 2013. ISSN 2218-5836. doi:10.5312/wjo.v4.i2.32. URL http://www.wjgnet.com/2218-5836/full/v4/i2/32.htm. → page 1

[17] I. Hacihaliloglu. Ultrasound imaging and segmentation of bone surfaces: A review. *TECHNOLOGY*, 05(02):74–80, jun 2017. ISSN 2339-5478. doi:10.1142/S2339547817300049. URL http://www.worldscientific.com/doi/abs/10.1142/S2339547817300049. → page 12

[18] K. Hara. 3D ResNets for Action Recognition, 2018. URL https://github.com/kenshohara/3D-ResNets-PyTorch. → pages 49, 50, 76, 86

[19] K. Hara, H. Kataoka, and Y. Satoh. Learning spatio-Temporal features with 3D residual networks for action recognition. *Proceedings - 2017 IEEE International Conference on Computer Vision Workshops, ICCVW 2017*, 2018-January:3154–3160, 2018. doi:10.1109/ICCVW.2017.373. → page 17

[20] K. Hara, H. Kataoka, and Y. Satoh. Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet? *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 6546–6555, 2018. ISSN 10636919. doi:10.1109/CVPR.2018.00685. → pages 17, 49, 50, 76, 86

[21] A. R. Hareendranathan, M. Mabee, K. Punithakumar, M. Noga, and J. L. Jaremko. A technique for semiautomatic segmentation of echogenic structures in 3D ultrasound, applied to infant hip dysplasia. *International Journal of Computer Assisted Radiology and Surgery*, 11(1):31–42, 2016. ISSN 18616429. doi:10.1007/s11548-015-1239-5. → pages 9, 11

[22] A. R. Hareendranathan, D. Zonoobi, M. Mabee, D. Cobzas, K. Punithakumar, M. Noga, and J. L. Jaremko. Toward automatic diagnosis of hip dysplasia from 2D ultrasound. *Proceedings - International Symposium on Biomedical Imaging*, pages 982–985, 2017. ISSN 19458452. doi:10.1109/ISBI.2017.7950680. → pages 13, 34

[23] A. R. Hareendranathan, D. Zonoobi, M. Mabee, C. Diederichs, K. Punithakumar, M. Noga, and J. L. Jaremko. Semiautomatic classification of acetabular shape from three-dimensional ultrasound for diagnosis of infant hip dysplasia using geometric features. *International Journal of Computer Assisted Radiology and Surgery*, 12(3):439–447, 2017. ISSN 18616429. doi:10.1007/s11548-016-1510-4. → pages 11, 15

[24] W. H. Harris. Etiology of osteoarthritis of the hip. *Clinical orthopaedics and related research*, 213:20–33, dec 1986. ISSN 0009-921X. URL http://europepmc.org/abstract/MED/3780093. → page 1

[25] M. Harris-Hayes and N. K. Royer. Relationship of Acetabular Dysplasia and Femoroacetabular Impingement to Hip Osteoarthritis: A Focused Review. *PM&R*, 3(11):1055 – 1067.e1, 2011. ISSN 1934-1482. doi:https://doi.org/10.1016/j.pmrj.2011.08.533. URL http://www.sciencedirect.com/science/article/pii/S1934148211010768. → page 1

[26] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. → page 49

[27] F. T. Hoaglund. Primary Osteoarthritis of the Hip: A Genetic Disease Caused by European Genetic Variants. *JBJS*, 95(5), 2013. ISSN 0021-9355. URL https://journals.lww.com/jbjsjournal/Fulltext/2013/03060/Primary{_}Osteoarthritis{_}of{_}the{_}Hip{_}{_}A{_}Genetic.11.aspx. → page 7

[28] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017-January: 2261–2269, 2017. doi:10.1109/CVPR.2017.243. → page 49

[29] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *32nd International Conference on Machine Learning, ICML 2015*, 1:448–456, 2015. → page 27

[30] F. Isensee, J. Petersen, A. Klein, D. Zimmerer, P. F. Jaeger, S. Kohl, J. Wasserthal, G. Koehler, T. Norajitra, S. Wirkert, and K. H. Maier-Hein. nnu-net: Self-adapting framework for u-net-based medical image segmentation, 2018. → pages 50, 83

[31] J. Jackson, M. Runge, and N. Nye. Common Questions About Developmental Dysplasia of the Hip. *American Family Physician*, 90(12):

843–850, 2014. URL https://www.aafp.org/afp/2014/1215/p843.html. → page 1

[32] J. L. Jaremko, M. Mabee, V. G. Swami, L. Jamieson, K. Chow, and R. B. Thompson. Potential for Change in US Diagnosis of Hip Dysplasia Solely Caused by Changes in Probe Orientation: Patterns of Alpha-angle Variation Revealed by Using Three-dimensional US. *Radiology*, 273(3):870–8, 2014. ISSN 1527-1315. doi:10.1148/radiol.14140451. URL http://www.ncbi.nlm.nih.gov/pubmed/24964047. → pages 1, 8

[33] A. Jennings. MATLAB Central: Sphere Fit (least squares), 2013. URL https://www.mathworks.com/matlabcentral/fileexchange/34129-sphere-fit-least-squared. → page 48

[34] Y. Jing, Y. Yang, Z. Feng, J. Ye, Y. Yu, and M. Song. Neural Style Transfer: A Review. *IEEE Transactions on Visualization and Computer Graphics*, pages 1–1, 2019. → page 88

[35] A. Karamalis, W. Wein, T. Klein, and N. Navab. Ultrasound confidence maps using random walks. *Medical Image Analysis*, 16(6):1101–1112, aug 2012. ISSN 13618415. doi:10.1016/j.media.2012.07.005. URL http://dx.doi.org/10.1016/j.media.2012.07.005https://linkinghub.elsevier.com/retrieve/pii/S1361841512000977. → page 28

[36] R. Kikinis, S. D. Pieper, and K. G. Vosburgh. *3D Slicer: A Platform for Subject-Specific Image Analysis, Visualization, and Clinical Support*, pages 277–289. Springer New York, New York, NY, 2014. ISBN 978-1-4614-7657-3. doi:10.1007/978-1-4614-7657-3_19. URL https://doi.org/10.1007/978-1-4614-7657-3{_}19. → page 37

[37] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization, 2014. → pages 37, 78

[38] T. K. Koo and M. Y. Li. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, 15(2):155–163, 2016. ISSN 15563707. doi:10.1016/j.jcm.2016.02.012. → pages 62, 63

[39] P. Kovesi. Symmetry and Asymmetry from Local Phase. In *Tenth Australian Joint Conference of Artificial Intelligence*, volume 190, pages 2—-4, 1997. → page 13

[40] H. P. Lehmann, R. Hinton, P. Morello, and J. a. Santoli. Developmental Dysplasia of the Hip Practice Guideline: Technical Report. *Pediatrics*, 105 (4):e57—-e57, 2000. ISSN 0031-4005. doi:10.1542/peds.105.4.e57. URL https://pediatrics.aappublications.org/content/105/4/e57. → page 7

[41] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar. Focal loss for dense object detection. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. → page 81

[42] R. T. Loder and E. N. Skopelja. The Epidemiology and Demographics of Hip Dysplasia. *ISRN Orthopedics*, 2011:1–46, 2011. ISSN 2090-6161. doi:10.5402/2011/238607. → page 1

[43] J. Long, E. Shelhamer, and T. Darrell. Fully Convolutional Networks for Semantic Segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, 2015. ISBN 9781467369640. doi:10.1109/CVPR.2015.7298965. URL http://arxiv.org/abs/1503.06350. → page 12

[44] M. G. Mabee, A. R. Hareendranathan, R. B. Thompson, S. Dulai, and J. L. Jaremko. An index for diagnosing infant hip dysplasia using 3-D ultrasound: the acetabular contact angle. *Pediatric Radiology*, 46(7):1023–1031, 2016. ISSN 14321998. doi:10.1007/s00247-016-3552-8. URL http://dx.doi.org/10.1007/s00247-016-3552-8. → page 9

[45] K. McHale and D. Corbett. Parental noncompliance with pavlik harness treatment of infantile hip problems. *Journal of pediatric orthopedics*, 9(6): 649—652, 1989. ISSN 0271-6798. doi:10.1097/01241398-198911000-00003. URL https://doi.org/10.1097/01241398-198911000-00003. → page 7

[46] F. Milletari, N. Navab, and S. A. Ahmadi. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. *Proceedings - 2016 4th International Conference on 3D Vision, 3DV 2016*, pages 565–571, 2016. doi:10.1109/3DV.2016.79. → page 36

[47] C. Morin, H. T. Harcke, and G. D. MacEwen. The infant hip: real-time us assessment of acetabular development. *Radiology*, 157(3):673–677, 1985. doi:10.1148/radiology.157.3.3903854. URL https://doi.org/10.1148/radiology.157.3.3903854. PMID: 3903854. → pages 2, 9, 57

[48] E. Mostofi, B. Chahal, D. Zonoobi, A. Hareendranathan, K. P. Roshandeh, S. K. Dulai, and J. L. Jaremko. Reliability of 2D and 3D ultrasound for infant hip dysplasia in the hands of novice users. *European Radiology*, 29 (3):1489–1495, mar 2019. ISSN 0938-7994. doi:10.1007/s00330-018-5699-1. URL http://link.springer.com/10.1007/s00330-018-5699-1. → pages 1, 9

[49] M. L. Murnaghan, R. H. Browne, D. J. Sucato, and J. Birch. Femoral nerve palsy in pavlik harness treatment for developmental dysplasia of the hip. *Journal of Bone and Joint Surgery - Series A*, 93(5):493–499, 2011. ISSN 15351386. doi:10.2106/JBJS.J.01210. → page 7

[50] S. Nakamura, S. Ninomiya, and T. Nakamura. Primary osteoarthritis of the hip joint in Japan. *Clinical orthopaedics and related research*, 241:190–196, apr 1989. ISSN 0009-921X. URL http://europepmc.org/abstract/MED/2924462. → pages 1, 7

[51] H. Ömeroğlu, N. Köse, and A. Akceylan. Success of Pavlik Harness Treatment Decreases in Patients 4 Months and in Ultrasonographically Dislocated Hips in Developmental Dysplasia of the Hip. *Clinical Orthopaedics and Related Research*, 474(5):1146–1152, 2016. ISSN 15281132. doi:10.1007/s11999-015-4388-5. → page 6

[52] Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. V. Dillon, B. Lakshminarayanan, and J. Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift, 2019. → pages 89, 90

[53] P. Pandey. *Real-time ultrasound bone segmentation and robust US-CT registration for surgical navigation of pelvic fractures*. PhD thesis, University of British Columbia, 2018. URL https://open.library.ubc.ca/cIRcle/collections/ubctheses/24/items/1.0375839. → page 12

[54] P. Pandey, P. Guy, A. Hodgson, and R. Garbi. Shadow Peak: Accurate Real-time Bone Segmentation for Ultrasound and Developmental Dysplasia of the Hip. In *19th Annual Meeting of the International Society for Computer Assisted Orthopaedic Surgery*, New York, 2019. → pages 13, 17, 26, 39

[55] G. Papandreou, L. C. Chen, K. P. Murphy, and A. L. Yuille. Weakly-and semi-supervised learning of a deep convolutional network for semantic

image segmentation. *Proceedings of the IEEE International Conference on Computer Vision*, 2015 Inter:1742–1750, 2015. → page 88

[56] O. Paserin. *Fully Automatic 3D Ultrasound Techniques for Improving Diagnosis of Developmental Dysplasia of the Hip in Pediatric Patients : Classifying Scan Adequacy and Quantifying Dynamic Assessment*. PhD thesis, The University of British Columbia, 2018. → pages xix, 14, 15, 16, 17, 18, 70, 73, 74, 76, 80, 85, 86, 115

[57] O. Paserin, K. Mulpuri, A. Cooper, and A. J. Hodgson. Automatic Near Real-Time Evaluation of 3D Ultrasound Scan Adequacy for Developmental Dysplasia of the Hip. *Computer Assisted and Robotic Endoscopy and Clinical Image-Based Procedures*, 10550:124–132, 2017. doi:10.1007/978-3-319-67543-5. URL http://link.springer.com/10.1007/978-3-319-67543-5. → page 14

[58] O. Paserin, K. Mulpuri, A. Cooper, A. J. Hodgson, and R. Garbi. Real Time RNN Based 3D Ultrasound Scan Adequacy for Developmental Dysplasia of the Hip. In *MICCAI 2018*, volume 8151, pages 365–373. Springer International Publishing, 2018. ISBN 978-3-642-40810-6. doi:10.1007/978-3-030-00928-1_42. URL http://link.springer.com/10.1007/978-3-642-40760-4http://link.springer.com/10.1007/978-3-030-00928-1{_}42. → pages xix, 14, 15, 16, 17, 29, 30, 70, 73, 74, 76, 79, 80, 86, 115

[59] O. Paserin, K. Mulpuri, A. Cooper, A. J. Hodgson, and R. Garbi. Automated dynamic 3d ultrasound assessment of developmental dysplasia of the infant hip. In *International Workshop on Computational Methods and Clinical Applications in Musculoskeletal Imaging*, pages 136–145. Springer, 2018. → pages 21, 57

[60] C. T. Price and B. A. Ramo. Prevention of Hip Dysplasia in Children and Adults. *Orthopedic Clinics*, 43(3):269–279, jul 2012. ISSN 0030-5898. doi:10.1016/j.ocl.2012.05.001. URL https://doi.org/10.1016/j.ocl.2012.05.001. → page 7

[61] N. Quader. *Automatic Characterization of Developmental Dysplasia of the Hip in Infants using Ultrasound Imaging*. PhD thesis, University of British Columbia, 2018. → pages xvi, 7, 8, 9, 11, 12, 14, 15, 16, 17, 18, 26, 39, 45, 53, 55, 58, 61, 62, 63, 69, 83, 84, 85, 111

[62] N. Quader, A. Hodgson, and R. Abugharbieh. Confidence Weighted Local Phase Features for Robust Bone Surface Segmentation in Ultrasound. In

CLIP 2014, volume 9958, pages 76–83, 2014. ISBN 978-3-319-46471-8. doi:10.1007/978-3-319-13909-8_10. URL http://link.springer.com/10.1007/978-3-319-46472-5http://link.springer.com/10.1007/978-3-319-13909-8{_}10. → pages 11, 12, 37, 39, 83

[63] N. Quader, A. Hodgson, K. Mulpuri, A. Cooper, and R. Abugharbieh. Towards Reliable Automatic Characterization of Neonatal Hip Dysplasia from 3D Ultrasound Images. In *MICCAI*, volume 9900, pages 602–609, 2016. ISBN 978-3-319-46725-2. doi:10.1007/978-3-319-46720-7_70. URL http://link.springer.com/10.1007/978-3-319-46726-9http://link.springer.com/10.1007/978-3-319-46720-7{_}70. → pages xvii, xviii, 9, 16, 17, 57, 61, 62, 63, 65, 67, 69, 85

[64] N. Quader, A. J. Hodgson, K. Mulpuri, A. Cooper, and R. Abugharbieh. A 3D Femoral Head Coverage Metric for Enhanced Reliability in Diagnosing Hip Dysplasia. In *MICCAI*, volume 10433, pages 100–107, 2017. ISBN 978-3-319-66181-0. doi:10.1007/978-3-319-66182-7_12. URL http://link.springer.com/10.1007/978-3-319-66182-7http://link.springer.com/10.1007/978-3-319-66182-7{_}12. → pages xvi, xvii, 9, 11, 16, 17, 18, 26, 45, 51, 53, 55, 57, 62, 63, 66, 69, 84, 85

[65] N. Quader, A. J. Hodgson, K. Mulpuri, E. Schaeffer, and R. Abugharbieh. Automatic Evaluation of Scan Adequacy and Dysplasia Metrics in 2-D Ultrasound Images of the Neonatal Hip. *Ultrasound in Medicine and Biology*, 43(6):1252–1262, 2017. ISSN 1879291X. doi:10.1016/j.ultrasmedbio.2017.01.012. → page 14

[66] N. Quader, E. K. Schaeffer, A. J. Hodgson, R. Abugharbieh, and K. Mulpuri. A Systematic Review and Meta-analysis on the Reproducibility of Ultrasound-based Metrics for Assessing Developmental Dysplasia of the Hip. *Journal of Pediatric Orthopaedics*, 38(6):e305–e311, 2018. ISSN 15392570. doi:10.1097/BPO.0000000000001179. → pages 1, 7, 8, 9, 57

[67] O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *MICCAI*, pages 234–241, 2015. ISBN 9783319245737. doi:10.1007/978-3-319-24574-4_28. URL http://link.springer.com/10.1007/978-3-319-24574-4{_}28. → pages 12, 17, 26, 27, 36, 37, 39, 83

[68] J. A. Rosenthal, X. Lu, and P. Cram. Availability of Consumer Prices From US Hospitals for a Common Surgical Procedure. *JAMA Internal Medicine*,

173(6):427–432, 2013. ISSN 2168-6106. doi:10.1001/jamainternmed.2013.460. URL https://doi.org/10.1001/jamainternmed.2013.460. → page 7

[69] F. Saberi Hosnijeh, M. E. Zuiderwijk, M. Versteeg, H. T. W. Smeele, A. Hofman, A. G. Uitterlinden, R. Agricola, E. H. G. Oei, J. H. Waarsing, S. M. Bierma-Zeinstra, and J. B. J. van Meurs. Cam Deformity and Acetabular Dysplasia as Risk Factors for Hip Osteoarthritis. *Arthritis & Rheumatology*, 69(1):86–93, 2017. doi:10.1002/art.39929. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/art.39929. → page 1

[70] M. D. Sewell, K. Rosendahl, and D. M. Eastwood. Developmental dysplasia of the hip. *BMJ*, 339(nov24 2):b4454–b4454, nov 2009. ISSN 0959-8138. doi:10.1136/bmj.b4454. URL http://www.bmj.com/cgi/doi/10.1136/bmj.b4454. → page 1

[71] D. Shorter, T. Hong, and D. A. Osborn. Cochrane Review: Screening programmes for developmental dysplasia of the hip in newborn infants. *Evidence-Based Child Health: A Cochrane Review Journal*, 8(1):11–54, 2013. ISSN 15576272. doi:10.1002/ebch.1891. URL http://doi.wiley.com/10.1002/ebch.1891. → page 1

[72] G. Tulder. Elastic deformations for N-dimensional images (Python, SciPy, NumPy, TensorFlow), 2018. URL https://github.com/gvtulder/elasticdeform. → page 38

[73] A. Valada, J. Vertens, A. Dhall, and W. Burgard. AdapNet: Adaptive semantic segmentation in adverse environmental conditions. In *Proceedings - IEEE International Conference on Robotics and Automation*, pages 4644–4651. IEEE, 2017. ISBN 9781509046331. doi:10.1109/ICRA.2017.7989540. → page 12

[74] 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*, 13(11):1707–1716, 2018. ISSN 18616429. doi:10.1007/s11548-018-1856-x. URL https://doi.org/10.1007/s11548-018-1856-x. → pages 12, 28, 29, 31, 34, 39

[75] M. Wang and W. Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153, 2018. → page 88

[76] P. Wang, V. M. Patel, and I. Hacihaliloglu. Simultaneous Segmentation and Classification of Bone Surfaces from Ultrasound Using a Multi-feature Guided CNN. In *MICCAI*, volume 9901, pages 134–142. Springer International Publishing, 2018. ISBN 978-3-319-46722-1. doi:10.1007/978-3-030-00937-3_16. URL http://link.springer.com/10.1007/978-3-319-46723-8http://link.springer.com/10.1007/978-3-030-00937-3{_}16. → pages 12, 28, 34

[77] G. Wilson and D. J. Cook. A survey of unsupervised deep domain adaptation, 2018. → page 88

[78] T. Woodacre, A. Dhadwal, T. Ball, C. Edwards, and P. J. Cox. The costs of late detection of developmental dysplasia of the hip. *Journal of Children's Orthopaedics*, 8(4):325–332, 2014. ISSN 18632548. doi:10.1007/s11832-014-0599-7. → pages 1, 7

[79] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017-January:5987–5995, 2017. doi:10.1109/CVPR.2017.634. → page 50

[80] S. Zagoruyko and N. Komodakis. Wide residual networks. *Procedings of the British Machine Vision Conference 2016*, 2016. doi:10.5244/c.30.87. URL http://dx.doi.org/10.5244/C.30.87. → page 49

[81] Z. Zhang, M. Tang, D. Cobzas, D. Zonoobi, M. Jagersand, and J. L. Jaremko. End-to-end detection-segmentation network with ROI convolution. *Proceedings - International Symposium on Biomedical Imaging*, 2018-April (Isbi):1509–1512, 2018. ISSN 19458452. doi:10.1109/ISBI.2018.8363859. → pages 13, 35

[82] D. Zonoobi, E. Mostofi, M. Mabee, and S. Pasha. Developmental Hip Dysplasia Diagnosis at Three-dimensional US: A Multicenter Study. *Radiology*, 287(3):1003–1015, 2018. ISSN 15271315. doi:10.1148/radiol.2018172592. → pages 9, 11, 15, 17, 18, 57, 58, 62, 63, 69, 85

Appendix A

§2 Supporting Materials

Automatic Characterization of the Neonatal Hip with 3-Dimensional and Tracked Ultrasound

Date form completed: __ __/__ __ __/ __ __ __ __ (Day / Month / Year) **Study ID: 3DUS19__ __ __**

Data Collection Form

Participant Demographics:

Date of Appointment: __ __/__ __ __/ __ __ __ __ (Day / Month / Year)

Chronologic age: _____ weeks (rounded to nearest whole number)

Gender: M F

Affected Hip: R L Bilateral

Familial History of DDH: Yes No

If yes, whom: _____

First Born Child: Yes No

Breech Presentation: Yes No

Caesarian Section: Yes No

Data Collection Form March 1, 2019

Figure A.1: Data collection form used in the clinical study

Appendix B

§4.1 Supporting Materials

Table B.1: Mean performance metrics for the four contrasted methods on a test set of 52 volumes from 13 participants.

	CSPS	CSPS_DDH	U-Net	3D-U-Net		ANOVA p-value
Precision	0.11	0.28	0.84	0.83		5.0E-115
Recall	0.83	0.28	1.00	0.87		7.9E-103
J	0.11	0.15	0.84	0.74		2.9E-127
DSC	0.19	0.26	0.91	0.85		1.5E-126
MED_R2P(mm)	0.15	3.54	0.00	0.12		3.3E-24
MED_P2R(mm)	14.52	9.16	0.93	0.29		1.4E-73
MED_max(mm)	14.52	9.23	0.93	0.31		2.6E-74
HD_R2P(mm)	1.65	11.63	0.39	3.38		3.2E-40
HD_P2R(mm)	33.57	32.78	16.50	3.56		3.8E-96
HD_max(mm)	33.57	32.78	16.50	5.46		1.5E-86
RMS_P2R(mm)	17.15	12.65	2.91	0.70		8.7E-84
CAI	0.35	0.31	0.94	0.92		3.1E-101
CDI	0.00	0.00	0.24	0.76		1.4E-56

Table B.2: Precision post hoc t-test p-values.

	CSPS	CSPS_DDH	U-Net	3D-U-Net
CSPS	1.0E+00	1.5E-16	7.6E-66	2.5E-76
CSPS_DDH	1.5E-16	1.0E+00	1.3E-46	2.0E-51
U-Net	7.6E-66	1.3E-46	1.0E+00	4.9E-01
3D-U-Net	2.5E-76	2.0E-51	4.9E-01	1.0E+00

Table B.3: Recall post hoc t-test p-values.

	CSPS	CSPS_DDH	U-Net	3D-U-Net
CSPS	1.0E+00	9.5E-43	7.0E-19	3.8E-02
CSPS_DDH	9.5E-43	1.0E+00	3.6E-62	2.1E-51
U-Net	7.0E-19	3.6E-62	1.0E+00	4.4E-30
3D-U-Net	3.8E-02	2.1E-51	4.4E-30	1.0E+00

Table B.4: Jaccard Coefficient post hoc t-test p-values.

	CSPS	CSPS_DDH	U-Net	3D-U-Net
CSPS	1.00E+00	2.55E-04	5.99E-66	2.80E-70
CSPS_DDH	2.55E-04	1.00E+00	4.62E-62	5.83E-65
U-Net	5.99E-66	4.62E-62	1.00E+00	2.45E-07
3D-U-Net	2.80E-70	5.83E-65	2.45E-07	1.00E+00

Table B.5: Dice-Sorensen Coefficient post hoc t-test p-values.

	CSPS	CSPS_DDH	U-Net	3D-U-Net
CSPS	1.0E+00	4.6E-04	1.9E-69	6.5E-69
CSPS_DDH	4.6E-04	1.0E+00	7.7E-62	2.3E-60
U-Net	1.9E-69	7.7E-62	1.0E+00	1.3E-06
3D-U-Net	6.5E-69	2.3E-60	1.3E-06	1.0E+00

Table B.6: MED_{R2P} post hoc t-test p-values.

	CSPS	CSPS_DDH	U-Net	3D-U-Net
CSPS	1.0E+00	3.3E-10	6.2E-15	2.5E-01
CSPS_DDH	3.3E-10	1.0E+00	8.0E-11	2.6E-10
U-Net	6.2E-15	8.0E-11	1.0E+00	7.2E-18
3D-U-Net	2.5E-01	2.6E-10	7.2E-18	1.0E+00

Table B.7: MED_{P2R} post hoc t-test p-values.

	CSPS	CSPS_DDH	U-Net	3D-U-Net
CSPS	1.0E+00	2.3E-10	3.2E-48	1.9E-51
CSPS_DDH	2.3E-10	1.0E+00	2.2E-23	3.7E-26
U-Net	3.2E-48	2.2E-23	1.0E+00	2.3E-04
3D-U-Net	1.9E-51	3.7E-26	2.3E-04	1.0E+00

Table B.8: MED_{max} post hoc t-test p-values.

	CSPS	CSPS_DDH	U-Net	3D-U-Net
CSPS	1.0E+00	2.5E-10	3.2E-48	2.1E-51
CSPS_DDH	2.5E-10	1.0E+00	4.8E-24	8.6E-27
U-Net	3.2E-48	4.8E-24	1.0E+00	3.5E-04
3D-U-Net	2.1E-51	8.6E-27	3.5E-04	1.0E+00

Table B.9: HD_{R2P} post hoc t-test p-values.

	CSPS	CSPS_DDH	U-Net	3D-U-Net
CSPS	1.0E+00	3.5E-23	7.7E-25	7.7E-03
CSPS_DDH	3.5E-23	1.0E+00	1.3E-26	4.4E-13
U-Net	7.7E-25	1.3E-26	1.0E+00	8.2E-06
3D-U-Net	7.7E-03	4.4E-13	8.2E-06	1.0E+00

Table B.10: HD_{P2R} post hoc t-test p-values.

	CSPS	CSPS_DDH	U-Net	3D-U-Net
CSPS	1.0E+00	1.7E-01	2.4E-33	3.1E-61
CSPS_DDH	1.7E-01	1.0E+00	4.7E-32	1.3E-60
U-Net	2.4E-33	4.7E-32	1.0E+00	1.1E-21
3D-U-Net	3.1E-61	1.3E-60	1.1E-21	1.0E+00

Table B.11: HD_{max} post hoc t-test p-values.

	CSPS	CSPS_DDH	U-Net	3D-U-Net
CSPS	1.0E+00	1.7E-01	2.4E-33	2.5E-52
CSPS_DDH	1.7E-01	1.0E+00	4.7E-32	1.5E-51
U-Net	2.4E-33	4.7E-32	1.0E+00	8.5E-16
3D-U-Net	2.5E-52	1.5E-51	8.5E-16	1.0E+00

Table B.12: CAI post hoc t-test p-values.

	CSPS	CSPS_DDH	U-Net	3D-U-Net
CSPS	1.0E+00	1.3E-01	3.8E-50	1.2E-49
CSPS_DDH	1.3E-01	1.0E+00	2.3E-55	4.5E-55
U-Net	3.8E-50	2.3E-55	1.0E+00	5.8E-02
3D-U-Net	1.2E-49	4.5E-55	5.8E-02	1.0E+00

Table B.13: CDI post hoc t-test p-values.

	CSPS	CSPS_DDH	U-Net	3D-U-Net
CSPS	1.0E+00	1.0E-04	2.8E-09	1.6E-36
CSPS_DDH	1.0E-04	1.0E+00	3.3E-09	1.8E-36
U-Net	2.8E-09	3.3E-09	1.0E+00	2.6E-16
3D-U-Net	1.6E-36	1.8E-36	2.6E-16	1.0E+00

Appendix C

§4.2 Supporting Materials

Table C.1: Results comparing the two proposed methods with the state-of-the art RFC for predicting the location of the femoral head. Note that the RFC and 3D-ResNet-50 were compared against the full sphere label as ground truth (as described in §4.2.1), whereas 3D-U-Net was compared against the semi-sphere cropped by bounding box B as ground truth.

	RFC	3D-ResNet-50_Reg	3D-U-Net_Seg		ANOVA p-value
Precision	0.46	0.62	0.73		1.1E-12
Recall	0.49	0.81	0.82		7.3E-20
J	0.29	0.53	0.62		4.4E-27
DSC	0.43	0.69	0.76		3.3E-26
CAE_x(mm)	1.63	1.61	1.04		9.3E-03
CAE_y(mm)	1.87	2.31	0.45		1.3E-11
CAE_z(mm)	2.27	1.19	0.66		2.2E-09
CED(mm)	3.90	3.35	1.42		5.3E-15
RAE(mm)	2.01	1.01	0.46		1.9E-17

Table C.2: Precision post hoc t-test p-values.

	RFC	3D-ResNet-50_Reg	3D-U-Net_Seg
RFC	1.0E+00	2.0E-05	6.1E-13
3D-ResNet-50_Reg	2.0E-05	1.0E+00	6.1E-04
3D-U-Net_Seg	6.1E-13	6.1E-04	1.0E+00

Table C.3: Recall post hoc t-test p-values.

	RFC	3D-ResNet-50_Reg	3D-U-Net_Seg
RFC	1.0E+00	4.0E-13	5.3E-13
3D-ResNet-50_Reg	4.0E-13	1.0E+00	5.1E-01
3D-U-Net_Seg	5.3E-13	5.1E-01	1.0E+00

Table C.4: Jaccard Coefficient post hoc t-test p-values.

	RFC	3D-ResNet-50_Reg	3D-U-Net_Seg
RFC	1.0E+00	3.6E-15	3.2E-24
3D-ResNet-50_Reg	3.6E-15	1.0E+00	5.3E-04
3D-U-Net_Seg	3.2E-24	5.3E-04	1.0E+00

Table C.5: DSC post hoc t-test p-values.

	RFC	3D-ResNet-50_Reg	3D-U-Net_Seg
RFC	1.0E+00	3.8E-14	4.2E-21
3D-ResNet-50_Reg	3.8E-14	1.0E+00	4.4E-04
3D-U-Net_Seg	4.2E-21	4.4E-04	1.0E+00

Table C.6: CAE_x post hoc t-test p-values.

	RFC	3D-ResNet-50_Reg	3D-U-Net_Seg
RFC	1.0E+00	9.4E-01	7.9E-03
3D-ResNet-50_Reg	9.4E-01	1.0E+00	3.6E-03
3D-U-Net_Seg	7.9E-03	3.6E-03	1.0E+00

Table C.7: CAE_y post hoc t-test p-values.

	RFC	3D-ResNet-50_Reg	3D-U-Net_Seg
RFC	1.0E+00	1.5E-01	8.1E-08
3D-ResNet-50_Reg	1.5E-01	1.0E+00	7.3E-16
3D-U-Net_Seg	8.1E-08	7.3E-16	1.0E+00

Table C.8: CAE_z post hoc t-test p-values.

	RFC	3D-ResNet-50_Reg	3D-U-Net_Seg
RFC	1.0E+00	3.2E-04	3.7E-08
3D-ResNet-50_Reg	3.2E-04	1.0E+00	2.8E-04
3D-U-Net_Seg	3.7E-08	2.8E-04	1.0E+00

Table C.9: CED post hoc t-test p-values.

	RFC	3D-ResNet-50_Reg	3D-U-Net_Seg
RFC	1.0E+00	9.7E-02	7.8E-13
3D-ResNet-50_Reg	9.7E-02	1.0E+00	2.3E-14
3D-U-Net_Seg	7.8E-13	2.3E-14	1.0E+00

Table C.10: RAE post hoc t-test p-values.

	RFC	3D-ResNet-50_Reg	3D-U-Net_Seg
RFC	1.0E+00	4.5E-07	9.1E-16
3D-ResNet-50_Reg	4.5E-07	1.0E+00	3.7E-06
3D-U-Net_Seg	9.1E-16	3.7E-06	1.0E+00

Appendix D

§4.3 Supporting Materials

Table D.1: Comparing the SD for paired inter-exam measures for the different DDH metrics (n=42 hips)

		SD
α_{3D} (°)	Quader[61]	2.6
	Ours	2.1
FHC_{3D} (%)	Quader[61]	2.9
	Ours	3.5
OCR (mm)	Quader[61]	-
	Ours	0.41

Appendix E

§5 Supporting Materials

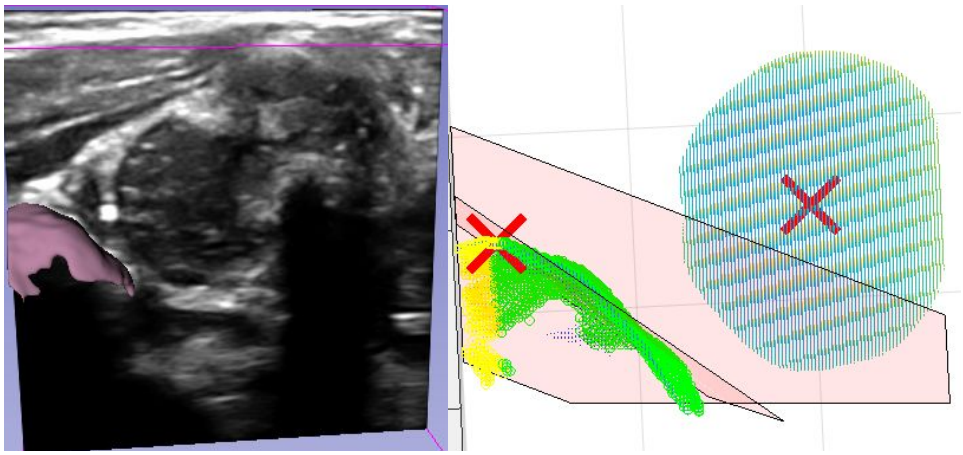


Figure E.1: An example case for which the ground truth label is "inadequate", our models predicted as "inadequate", but that the RNN predicted as "adequate". **Left:** the coronal view near the standard plane, with the 3D-U-Net pelvis bone surface prediction overlaid in pink. **Right:** the ilium and acetabulum point clouds after processing with the metrics extraction algorithm described in §4.3. We get a clear picture from these views that the probe is positioned too inferiorly, and that much of the ilium surface is not imaged. As a result, the bony rim appears to be misidentified, and the ilium plane appears to be incorrectly fitted, ultimately resulting in invalid α_{3D} and FHC_{3D} measurements.

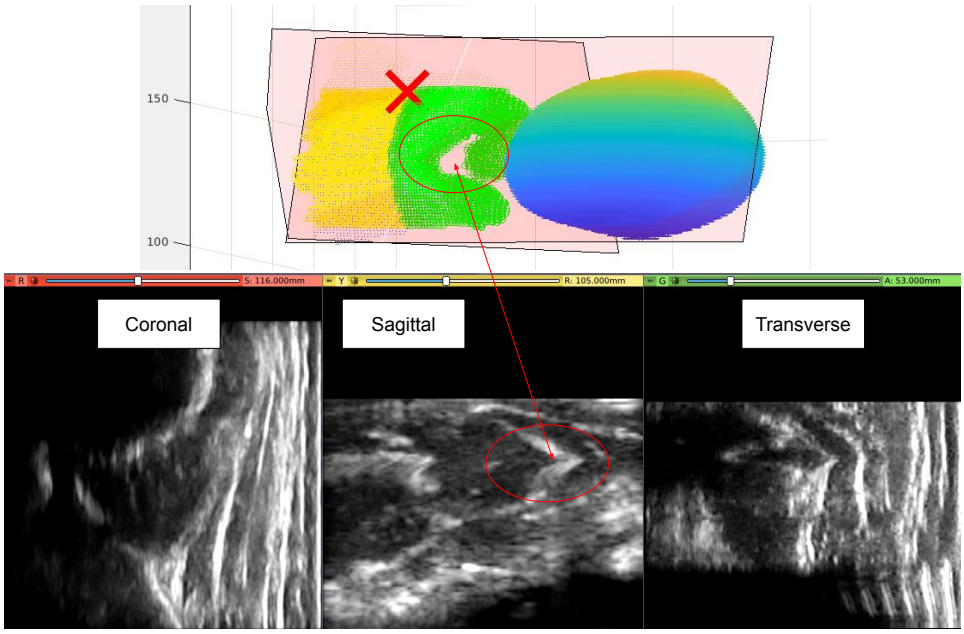


Figure E.2: Another example case for which the ground truth is "inadequate", our models predicted as "inadequate", but that the RNN predicted as "adequate". In this case we show the segmented points clouds in the **top** row, and the 3 anatomical planes in the **bottom** row. We can see in the sagittal view that there is a "smudge" due to movement artifact, circled in red. The effect of this on the acetabulum point cloud can be seen as a gap in the acetabulum that is usually not present in high-quality, adequate volumes.

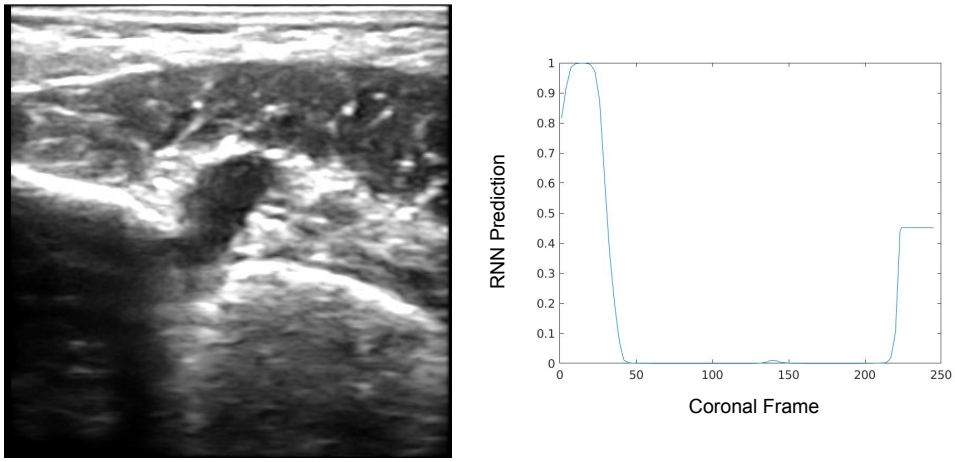


Figure E.3: Another example case for which the ground truth is "inadequate", our models predicted as "inadequate", but that the RNN predicted as "adequate". **Left:** we show our best attempt at locating the standard plane by browsing all the coronal slices. **Right:** the per-frame prediction of the RNN, from which the finall RNN prediction is made by thresholding and summing. We clearly see that the RNN incorrectly predicts very high scores for the first 40 slices, although none of these meet the criteria defined by Paserin [56, 58].