Imperial College London
Department of Computing

Computer-Assisted Intraoperative Ultrasound in Brain Tumour Surgery

Alistair Guy Weld

Submitted in part fulfillment of the requirements for the degree of
Doctor of Philosophy of Imperial College London
November 7, 2025

Copyright Declaration

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-Non Commercial-No Derivatives 4.0 International Licence (CC BY-NC-ND).

Under this licence, you may copy and redistribute the material in any medium or format on the condition that; you credit the author, do not use it for commercial purposes and do not distribute modified versions of the work.

When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text.

Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

I certify that, to the best of my knowledge, the work presented in this thesis is original, where correct and proper acknowledgement is made when required.

ALISTAIR GUY WELD

Date: November 7, 2025

"But I erred because I saw that I'd get nowhere on the straight path, and that to go crookedly was straighter."

Dead Souls - Nikolai Gogol

Dedication

To my mother and father, Charin and John Weld, and my other half, Beate Müller.

Acknowledgements

For your company and support during my PhD, I would like to thank - Supervisor: Stamatia Giannarou. Collaborators: Michael Dyck, Luke Dixon, Giulio Anichini, Giovanni Faoro, Julian Klodmann, Alin Albu-Schäffer and Sophie Camp. PhD colleagues: João Cartucho, Samyakh Tukra, Chi Xu, Joseph Davids, Alfie Roddan, Haozheng Xu, and Yihang Zhou.

Funding -

This research was carried out with support from the UK Research and Innovation Centre (UKRI) for Doctoral Training in AI for Healthcare (EP/S023283/1).

Abstract

Intraoperative imaging has been shown to be a vital tool for surgical decision making, particularly in the case of brain tumour surgery. Real-time acquisition of subsurface information is essential for achieving maximal safe resection, removing as much pathological tissue as possible while minimising damage to normal and critical tissue. Ultrasound has shown considerable promise in becoming the principal tool for assisting neurosurgeons. However, despite this potential, due to the lack of standardisation of training and use, the variability between centres is high. This factor has, as a consequence, contributed to the lack of research on this imaging tool, especially compared to other similar modalities. Brain tumours are an especially complex type of tumour and the surgical process is likely to incur many confounders, making real-time subsurface imaging essential. In particular, in the case of diffuse tumours, where due to the inflitrative properties, the margins suffer from ill-definition, which can introduce significant uncertainty and subjectivity during surgical decision making.

This thesis presents research on computer-assisted methods for intraoperative ultrasound for brain tumour resection. The work was carried out as part of the Imperial–Technische Universität München (TUM) Joint Academy of Doctoral Studies, in collaboration with the Department of Neurosurgery and Imaging at Charing Cross Hospital. The work at Imperial, conducted at the Hamlyn Centre, primarily addressed computer vision topics relevant to the clinical context, while TUM and the Deutsches Zentrum für Luft- und Raumfahrt (German Aerospace Centre) focused on robotics and control systems.

The thesis contributions can be grouped into three categories: Pathological Tissue Delineation, Probe Manipulation Analysis, Resource Leveraging. Together, the work in these chapters aims to expand the understanding of ultrasound brain tumour imaging and to present solutions to enable real-time computer-assisted support for surgeons and facilitate automation through robotic systems.

List of Medical and Medical Imaging Abbreviations

List of Abbreviations

Abbreviation	Definition
AS	Acoustic Shadow
B-mode	Brightness-mode
CT	Computed Tomography
GTR	Gross Total Resection
HGG	High Grade Glioma
IO	IntraOperative
IOUS	IntraOperative UltraSound
iMRI	Intraoperative Magnetic Resonance Imaging
LGG	Low Grade Glioma
MIS	Minimally Invasive Surgery
MRI	Magnetic Resonance Imaging
NHS	National Health Service
RF	Radio-Frequency
STR	SubTotal Resection
US	Ultrasound

List of Miscellaneous Technical Abbreviations

List of Abbreviations

Abbreviation	**Definition**
3DV	3 Dimensional Vision
AN	Annotators
COCO	Common Objects in COntext (dataset)
DEF	DEFormable
DETR	DEtection TRansformer (neural network)
DICOM	Digital Imaging and Communications in Medicine (data format)
DORN	Deep Ordinal Regression Network (neural network)
KITTI	Karlsruhe Institute of Technology and Toyota Technological Institute (dataset)
MTL	Multi-Task Learning
PSMNet	Pyramid Stereo Matching Network (neural network)
RAFT	Recurrent All-Pairs Field Transforms (neural network)
ResNet	Residual Network (neural network)
RF	Radio-Frequency
ROC	Receiver Operating Characteristic (FPR, TPR)
RGB	Red Green Blue
SCARED	Stereo Correspondence and Reconstruction of Endoscopic Data (dataset)
STL	Single-Task Learning

List of Mathematical Definitions and Descriptions

List of Mathematical Definitions

Abbreviation	**Definition**	**Mathematical description**
ACC	Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$
ADAM	Adaptive Moment Estimation	$\theta_{t+1} = \theta_t - \alpha \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \varepsilon}$
AUC	Area Under the Curve (ROC)	$\sum_{i=1}^{n-1} (\mathrm{FPR}_{i+1} - \mathrm{FPR}_i) \cdot \frac{\mathrm{TPR}_{i+1} + \mathrm{TPR}_i}{2}$
BBox	Bounding Box	$(u_{\min},\, v_{\min},\, u_{\max},\, v_{\max})$
BCE	Binary Cross-Entropy	$-\frac{1}{n} \sum_{i=1}^{n} \big[y_i \log(p_i) + (1 - y_i) \log\big(1 - p_i\big) \big]$
Brier	Brier Score	$\frac{1}{n} \sum_{i=1}^{n} (p_i - y_i)^2$
DSC	Sørensen–Dice coefficient	$\frac{2\,TP}{2\,TP + FP + FN}$
FN	False Negative	$\sum_{i=1}^{n} [\, y_i = 1 \ \wedge \ \hat{y}_i = 0 \,]$
FP	False Positive	$\sum_{i=1}^{n} [\, y_i = 0 \ \wedge \ \hat{y}_i = 1 \,]$
FPR	False Positive Rate	$\frac{FP}{FP + TN}$
HD	Hausdorff Distance	$\max \left\{ \sup_{y \in Y} \inf_{\hat{y} \in \hat{Y}} \|y - \hat{y}\|,\ \sup_{\hat{y} \in \hat{Y}} \inf_{y \in Y} \|\hat{y} - y\| \right\}$
Huber	Huber Loss	$\frac{1}{n} \sum_{i=1}^{n} \begin{cases} \frac{1}{2} (\hat{y}_i - y_i)^2, & \text{if } \lvert\hat{y}_i - y_i\rvert < 1 \\ \lvert\hat{y}_i - y_i\rvert - \frac{1}{2}, & \text{if } \lvert\hat{y}_i - y_i\rvert \geq 1 \end{cases}$
I	Image	$I \in \mathbb{R}^{H \times W \times C}$
IoU	Intersection over Union	$\frac{\lvert y \cap \hat{y} \rvert}{\lvert y \cup \hat{y} \rvert}$
IQR	InterQuartile Range	$IQR = Q_3 - Q_1$
KDE	Kernel Density Estimation	$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h} \right)$

Continued on next page

Table 3: Mathematical definitions (continued)

Abbreviation	**Definition**	**Mathematical description**
MSE	Mean Squared Error	$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
PPV	Precision	$\frac{TP}{TP+FP}$
RMSE	Root Mean Square Error	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$
Seg	Segment / Mask	$S \in \{0,1\}^{H \times W}$
SGD	Stochastic Gradient Descent	$\theta_{t+1} = \theta_t - \alpha \nabla_\theta \mathcal{L}(\theta_t)$
TN	True Negative	$\sum_{i=1}^{n}[\, y_i = 0 \ \wedge \ \hat{y}_i = 0\,]$
TNR	True Negative Rate (Specificity)	$\frac{TN}{TN+FP}$
TP	True Positive	$\sum_{i=1}^{n}[\, y_i = 1 \ \wedge \ \hat{y}_i = 1\,]$
TPR	True Positive Rate (Recall)	$\frac{TP}{TP+FN}$
Tversky	Tversky Loss	$1 - \frac{\sum_{i=1}^{n} p_i y_i}{\sum_{i=1}^{n} p_i y_i + \alpha_{tv}\sum_{i=1}^{n} p_i(1-y_i) + \beta_{tv}\sum_{i=1}^{n} y_i(1-p_i)}$

Note: y_i denotes ground truth label, $\hat{y}_i$ predicted label; p_i predicted probability; $Y, \hat{Y}, P$ sets of points representing ground truth and predictions; x denotes the input to a system/process; n number of samples; Q_1, Q_3 lower and upper quartiles; $[H \times W \times C]$ denotes height, width and channel; $u_{\min}, v_{\min}, u_{\max}, v_{\max}$ image coordinates; S segmentation mask; θ neural network parameters/weights; α learning rate; $\mathcal{L}$ loss function; ε small constant for numerical stability; $\hat{m}_t, \hat{v}_t$ bias-corrected first and second moment estimates; K kernel function; h kernel bandwidth; $\cap$ and $\cup$ set intersection and union; $\|\cdot\|$ Euclidean norm; α_{tv}, β_{tv} Tversky parameters.

Contents

Copyright Declaration i

Dedication iv

Acknowledgements v

Abstract vi

1 Introduction 1

1.1 Motivations 1

1.1.1 Healthcare Benefits of Robotic and Computer-Assisted Surgery 2

1.2 Thesis Structure And Summary 4

1.3 Publications 9

1.3.1 Thesis Specific Publications 9

1.3.2 Other Publications 10

1.3.3 Presentations and Posters 12

1.4 Coronavirus Disease 2019 Impact Statement 14

1.5 Descriptions, Conventions and Short Handings 15

2 Literature Review **16**

2.1 Clinical Background . . . 17

2.1.1 Brain Tumours . . . 17

2.1.2 Brain Tumour Surgery . . . 19

2.1.3 Intraopeartive Imaging . . . 21

2.1.4 Ultrasound . . . 22

2.2 Technical Background - Computer Vision Algorithms For Limited Data Applications . . . 25

2.2.1 Non-learning Examples in Surgical Application . . . 26

2.2.2 Deep Learning With Limited Data . . . 27

2.2.2.1 Data-Focussed Examples . . . 28

2.2.2.2 Transfer learning . . . 30

2.2.2.3 Learning Without Ground Truth . . . 32

2.2.2.4 Multi-Task Learning . . . 33

2.2.2.5 Meta-Learning . . . 33

2.3 Technical Background - Ultrasound . . . 35

2.3.1 Ultrasound Hardware and Imaging . . . 35

2.3.2 Ultrasound Registration . . . 36

2.3.3 Ultrasound Segmentation . . . 36

2.3.4 Thesis Bridge . . . 37

I Pathological Tissue Delineation 39

3 Interobserver Variance Of Tumour Annotation 41

3.1 Introduction . . . 41

3.2 Materials and Methods . . . 44

3.2.1 Data . . . 44

3.2.2 Annotation protocol . . . 46

3.2.3 Statistical analysis . . . 46

3.2.4 Tumour Border Pixel Dispersion Substudy . . . 47

3.3 Results . . . 48

3.3.1 Inter-Neurosurgeon annotation variance . . . 48

3.3.2 Tumour Border Pixel Dispersion Substudy . . . 49

3.3.3 Neuroradiologist-Neurosurgeon annotation variance . . . 51

3.3.4 Improving segmentation accuracy using overlaid bounding boxes as a guide 54

3.4 Discussion . . . 56

3.4.1 Historical background and clinical landscape . . . 56

3.4.2 Findings, challenges, and future perspectives . . . 56

3.5 Conclusion . . . 61

4 Robust Fine-Tuning For Tumour Detection 62

4.1 Introduction . . . 62

4.2 Method and Experimental Setup . . . 66

4.2.1 Data . . . 66

4.2.2 Robust Fine-Tuning . . . 66

4.2.3 Training Framework and Comparisons Setup . . . 68
4.3 Results . . . 71
4.3.1 Detection Performance . . . 71
4.3.2 Inherent Robustness to Noise . . . 74
4.3.3 Attention maps . . . 74
4.4 Conclusion . . . 75

5 A Sparse Annotation Method Based On Confidence Grading 76

5.1 Introduction . . . 76
5.2 Methodology . . . 79
5.2.1 Annotation Protocol . . . 79
5.2.2 Confidence Map Construction . . . 80
5.3 Experiments And Results . . . 82
5.3.1 Confidence Map And Interobserver Variance Similarity . . . 82
5.3.1.1 Metric Similarity . . . 83
5.3.1.2 Similarity Across Threshold Levels . . . 84
5.3.2 Margin Uncertainty Correlation . . . 86
5.3.3 Downstream Application: Soft Label Learning . . . 87
5.4 Discussion . . . 89
5.5 Conclusion . . . 90

II Probe Manipulation Analysis 92

6 Probe-Tissue Contact Analysis 95

6.1 Introduction . . . 95
6.2 Methodology . . . 98
6.2.1 Detection of tissue features . . . 98
6.2.2 Artefact Detection and Confidence Map Generation . . . 100
6.3 Resources And Evaluation Setup . . . 101
6.3.1 Intraoperative Data . . . 101
6.3.1.1 Data curation . . . 101
6.3.1.2 Speckle noise robustness evaluation setup . . . 102
6.3.1.3 Classification evaluation and comparison setup . . . 103
6.3.1.4 Robotic Setup . . . 105
6.3.2 Medical Phantom Data . . . 108
6.3.2.1 Data curation . . . 108
6.3.2.2 Confidence map evaluation and comparison setup . . . 109
6.3.2.3 Classification generalisability . . . 109
6.4 Results and Discussion . . . 110
6.4.1 Noise Robustness . . . 111
6.4.2 Acoustic Shadow Classification . . . 111
6.4.2.1 Generalisability Study . . . 111
6.4.2.2 Parameter Perturbation Study . . . 112
6.4.3 Robotics results . . . 113
6.4.4 Confidence Map Behaviour . . . 115
6.5 Conclusion . . . 116

III Resource Leveraging **117**

7 Structured Light Regularising **120**

7.1 Introduction . . . 120

7.2 Related Work . . . 122

7.2.1 Active Sensors . . . 122

7.2.2 Depth estimation . . . 122

7.2.3 Multi Task Learning . . . 123

7.3 Method . . . 124

7.3.1 Dataset generation . . . 124

7.3.2 Structured light reconstruction . . . 128

7.3.3 Direct disparity estimation with multi task leaning . . . 130

7.4 Results . . . 133

7.4.1 Evaluation of structured light reconstruction . . . 133

7.4.2 Evaluation of the benefit of using multi task learning for the enhancement of disparity learning . . . 136

7.5 Conclusion . . . 141

8 Inexpensive Brain Model With Tumour Simulation **142**

8.1 Introduction . . . 142

8.2 Materials and Methods . . . 144

8.3 Results . . . 147

8.3.1 Phantom analysis . . . 147

8.3.2 Ex-Vivo and agar . . . 148

8.3.3 Necrosis Simulation . . . 150

8.4 Discussion . . . 152

8.5 Conclusion . . . 155

9 Conclusion **157**

9.1 Contributions . . . 158

9.1.1 Pathalogical Tissue Delineation . . . 158

9.1.2 Probe Manipulation Analysis . . . 159

9.1.3 Resource Leveraging . . . 160

9.2 Limitations and Future Work . . . 162

9.2.1 Craniotomy 3DV Study . . . 162

9.2.2 Ultrasound Machine Settings . . . 162

9.2.3 Multicentre studies . . . 163

Bibliography **163**

Appendix **200**

A A Background on Machine Learning Algorithms **201**

A.1 Early Developments . . . 202

A.2 Deep Learning . . . 205

B Standardisation of Convex Ultrasound **208**

B.1 Introduction . . . 208

B.2 Methodology . . . 212

B.2.1 Ultrasound Plane Masking . . . 213
B.2.2 Centre Of Plane Calculation . . . 215
B.2.3 Radial Boundary Detection . . . 217
B.2.4 Annulus Sector Parameters Calculation . . . 219
B.2.5 Scan line extraction and linearisation . . . 220
B.3 Experimental Setup and Results . . . 222
B.3.1 Data . . . 222
B.3.2 Key Point Detection Evaluation . . . 223
B.3.2.1 Controlled Experiment . . . 223
B.3.2.2 Public, Noisy Data . . . 224
B.3.3 Linearisation Study . . . 226
B.3.3.1 Deformation . . . 226
B.3.3.2 Inversion . . . 227
B.3.3.3 Radio-Frequency Data . . . 228
B.4 Conclusion . . . 230

List of Tables

3.1 Patient information. KPS = Karnofsky Performance Status, ASA = American Society of Anesthesiologists Grade, SCC = Squamous Cell Carcinoma, SS = Supratentorial Subependymoma, DNET = Dysembryoplastic Neuroepithelial Tumor, TB = Tuberculosis, Lt = Left, Rt = Right. 45

3.2 Shown are the IoU and DSC similarity results. The closer the value to 1 the more similar the annotations are. IoU is Intersection Over Union, DSC is Sørensen-Dice Similarity Coefficient. 52

3.3 Shown are the HD similarity results. The closer the value to 0 pixels the better aligned the annotation margins are. HD is Hausdorff Distance. 53

4.1 Detection performance for the models produced using unconstrained and constrained gradient descent. The epoch number corresponds to the best performing checkpoint on the test set. 72

4.2 Gaussian Noise. 73

4.3 Brightness & Contrast. 73

4.4 Speckle Noise. 73

4.5 Pixel Drop. 73

5.1 Training a UNet on soft and hard annotations with varying loss functions. The results correspond to the average across the 5-folds. For all loss functions, the soft trained networks achieves better test performance than the hard trained networks. 89

6.1 Scan line classification results. A positive sample corresponds to a scan line affected by AS. Highlighted in grey is the best score for each metric. The results have been averaged over the 3 folds. 110

6.2 The rankings (R#) of the methods using the F_β scores. The results are a product of averaging across the 3 folds. 110

6.3 The rankings (R#) of the methods' performances on the generalisability test on the probe pose analysis data, using the F_β scores. For all scans, *Topol* achieves the best rank. 110

6.4 The method's NRMSE results are shown for each scan. In all cases, *Topol* produces the best score. 111

6.5 Parameter perturbation, performance drift measured via coefficient of variation values (%). 115

7.1 Mean absolute error (MAE) produced by the proposed structured light network (with the disparity produced using the cross correlation) and the direct disparity network. 135

7.2 Comparison of multi task learning vs single task learning on the virtual dataset. Two types of experiments were run, the first was using the full training set, the second was using 1/16th of the training set. Assessing the general performance of each algorithm when training on a large and a small dataset. 138

7.3 Generalisability evaluation using SCARED [AMW+21] and the Middlebury 2014 [SHK+14]. The choice of displaying the depth error for SCARED and disparity error for Middlebury reflects what is commonly done. The results show that the *stl* model has achieved poorer generalisability than the MTL approaches. TS denotes the test sets in SCARED. IQR denotes the interquartile range 139

8.1 Cost, acquisition and setup 144

8.2 Qualitative assessment using a five-point Liekert scale (1 = poor to 5 = excellent) 147

8.3 Hypothetical training workshop itinerary utilising ex vivo models and agar solution 152

B.1 MSE and MAAD results from the public data evaluation. 225

List of Figures

1.1 The research space for robotic intraoperative ultrasound in brain surgery. Subdivided into four key areas (Robotics, Image Analysis, Resources, Clinical goals + Imaging) and their respective subtopics . The colours of the boxes indicate: red = Part 1; green = Part 2; blue = Part 3; grey = topics addressed during the PhD but not included in the thesis. 5

1.2 Thesis workflow and system logic for robotic IOUS in brain tumour surgery. The logic proceeds from methods that enable experimentation (Part III) to ultrasound-driven probe–tissue coupling (Part II), and finally to pathological tissue analysis within the acquired images (Part I). The box colours are the same as in Figure 1.1. 6

2.1 Brain tumour illustration - Wellcome Collection, United Kingdom - CC BY. https://www.europeana.eu/item/9200579/use8h2y8 17

2.2 Example images from the BRATS dataset showing different MRI sequences. Segmentation legend: blue is necrotic and non-enhancing tumor core, green is peritumoral edema, silver is Gadolinium enhancing tumour. 19

2.3 Illustration of a craniotomy process - "Drawing of the brain showing a subcortical clot." from Harvey Cushing Photograph Collection, Medical Historical Library, Cushing/Whitney Medical Library, Yale University Library Digital Collections, Image ID: 32325691, https://collections.library.yale.edu/catalog/15951324. 20

2.4 Simplified diagram of the US scan within a craniotomy. 22

2.5 Intraopertative ultrasound scan captured at Charing Cross Hospital. 23

3.1 Inter neurosurgeon annotation variance. From top to bottom - ID 14, ID 16, ID 18. ID is the image index. 48

3.2 The top row/example shows the effect on the boundary of a binary circle when Gaussian blurring is applied - using a kernel density estimate plot. The bottom row/example shows the same but using a box filter, where the circle also now has protrusions, unaffecting the defined boundary represented by the yellow mask. 49

3.3 Top is a plot of all tumour boundary pixel intensities distributions. Blue highlights unimodal distributions and red highlights other. Bottom is an example of tumour boundary is shown using ID 009. ID is the image index. 50

3.4 Box and Whisker plots of the annotators IoU, DSC and HD scores on the 30 images - red = segmentation, black = bounding box. What is highlighted is the greatly reduced median score and IQR when using the bounding box method. IoU is Intersection Over Union, DSC is Sørensen-Dice Similarity Coefficient, HD is Hausdorff Distance and IQR is Interquartile Range. 51

3.5 Uncertainty caused by unclear boundary. The left images are the original images, the middle from the neuroradiologist and the right from a neurosurgeon. From top to bottom - ID 13 An3, ID 07 An3, ID 18 An2, ID 19 An2. ID is the image index. 55

3.6 Examples for where the segments have large dissimilarity whilst the bounding boxes do not. The left images are the original images, the middle from the neuroradiologist and the right from a neurosurgeon. From top to bottom - ID 11 AN 2, the bottom from ID 26 AN 1. ID is the image index. 55

4.1 Sample ground truth annotations, where each row corresponds to a different scan (different patient) and going from left to right is going forward in time. 67

4.2 Example outputs from DETR on fold 3 test data. The red bounding box is the ground truth and the green is the best predicted. The brighter the yellow, the more attention. 71

4.3 Impact of constrained gradient descent on the final model weights. The solid lines show DETR and the dotted lines DEF DETR. Unconstrained gradient descent is highlighted in red while constrained gradient descent with blue. The circle represents the mean difference between final model and pre-trained parameters, while the lines highlight the standard deviation. The y axis is log scaled. 72

5.1 Tumour margin delineation uncertainty among four professional annotators. The three vectors show visible ambiguity over boundary point. 76

5.2 Sample ground truth annotations. 82

5.3 Mean Receiver Operator Characteristic (ROC) curves across threshold interobserver agreement levels. Left subfigure shows the ROC will all pixels included, while the right subfigure shows the ROC only using union area between the two annotations. 84

5.4 Metric performance across different threshold levels, where the thresholds are applied to both annotations. 85

5.5 The relationship between annotator variance and confidence annotation at the tumour margin (non-1 values) shows a clear linear trend, supported by correlating statistical evidence. 86

5.6 Test set output comparison from training on soft (left column) and hard (right column) ground truth. Where in each column, the left image is the ground truth and right the UNet output. 90

6.1 Ultrasound segmentation of IOUS images. The top row shows suboptimal US image capture, and the bottom row shows good US image capture. Annotations: green = solid tumour, orange = tumour necrotic core, yellow = cerebral sulci, dark blue = oedema, turquoise = surrounding normal brain, yellow = coupling fluid/probe brain interface. 96

6.2 Sample annotations of b-mode ultrasound images. The red line delineates acoustic shadowing secondary to poor contact whilst the blue line defines acoustic shadowing related to other causes (in this case echogenic blood clot in a resection cavity). 97

6.3 The image feedback at different frames from Scan 1 - abdominal. A normal (Gaussian) curve has been used to approximate the confidence behaviour of an ultrasound probe as it tilts, where the optimal imaging is produced when the probe is perpendicular to the tissue surface. This is shown in frame 50, which provides the highest quality visible feedback. As the probe moves away from perpendicularity, image quality reduces in a symmetric, Gaussian pattern. This is evident from the neighbouring frames. Frames 30 and 70, the furthest shown examples from optimal, show the lowest image quality and the least amount of visual feedback. Frames 40 and 60 produce images of slightly reduced quality than frame 50, particularly in terms of sharpness. 101

6.4 Sample simplicial complex outputs. These complexes indicate areas of significant spatial intensity variation within the ultrasound images. Areas where there are dense complexes, correlate with high confidence in tissue visibility; correlating with a confidence map output of higher perceptual quality. 105

6.5 Example of MIRO robotic arm setup at the DLR site in Oberpfaffenhofen. In this instance the experiment is being conducted on a porcine brain, for separate work as part of the collaboration. 106

6.6 Scanning on the home made phantom. 107

6.7 The FAST / acute abdomen and Kezlex brain phantom. 108

6.8 AS detection on two US images. The top row shows the annotation, the second is from $Baseline_1$, the third is from $VGG16$ and the bottom is from Topol. The orange areas are detected AS. 112

6.9 Graphs displaying the effect of individually varying the parameters on the performance of the whole system - using the unseen fold 3 data. blue lines = ACC, orange lines = TNR, green lines = TPR and red lines = PPV. 113

6.10 The per frame average confidence vector outputs - x axis = frames and y axis = average confidence. The blue lines show the method's outputs, *Topol* top, *Randomwalk* middle and *DAG* bottom row. The black dotted lines are the scaled normal curve. The dotted red lines are where the annotators perceived the best images. What can be observed is that the output of *Topol* follows the desired trajectory, closer than the other methods. 114

6.11 Robot base homogeneous transformation matrices and desired rotation and tilt commands. 114

6.12 Acoustic shadow behaviour and optimisation for robot scan. 115

7.1 Demonstration of the virtual environment. The virtual camera is represented in orange and the projector object is represented in yellow. 124

7.2 Sample objects used to create the virtual dataset. 125

7.3 Visualisation of the binary code structured light patterns, before binarizing. The images are from the left camera. 126

7.4 Shown in (a) and (b) is the setup for the medical dataset collection. Image (a) shows an example of tissue captured using the dVRK laparascope. (b) does not contain projection patterns and would be fed as an input to the neural network. 126

7.5 The medical dataset image capture setup. 127

7.6 Examples of ex vivo used. 128

7.7 The calibration of the laprascope using MATLAB's Stereo Camera Calibrator app. 128

7.8 Simple block diagram representation of the multi task learning framework that has been applied on top of the PSMNet architecture. 131

7.9 This figure contains an example output from the structured light network applied on the medical dataset. The top row contains the ground truth images and the bottom row shows the predictions. 134

7.10 Sample from virtual dataset validation. 135

7.11 Disparity accuracy and uncertainty correlation. The brighter the pixel, the greater the uncertainty and the less accurate. The ellipses are drawn over the same pixels in all three images. Correlating high energy is observed in the areas within the ellipses. 136

7.12 Log scaled correlation between uncertainty and prediction error. With a line of best fit drawn, in red, highlighting the positive correlation. The colourmap is set to 'jet'. The x and y axis are both normalised. 137

7.13 The blue dotted line is the performance of the *stl* model, which is used as a benchmark. What can be seen is that the only stably training algorithm is *unc*, in red, which also produces the best accuracy. All other models begin to overfit after 20 epochs. 139

8.1 UpSurgeOn brain box phantom synthetic material structure. The inner red material is the simulated tumour. 144

8.2 An ovine sample used to collect the data, before the agar solution was injected. The container has the dimensions 6.5cm x 7.0cm. 145

8.3 UpSurgeOn brain box phantom image from our machine. 147

8.4 Bovine sample. Left shows ultrasound feedback of phantom without tumour addition. Right shows ultrasound feedback with agar injected. 148

8.5 Ovine sample. B-mode image of sheep brain (A) with magnified image of gyrus (B) demonstrating discernible normal cerebral architecture with sulci and pia (outlined by green dotted line), cortex (outlined by orange dotted line) and surrounding white matter. B-mode image of the same sheep brain after localised infiltration of agar to mimic a hypercellular, echogenic tumour (C) with magnified image of tumour mimic demonstrating clear margins (D). 149

8.6 Human intraoperative ultrasound images of Glioblastoma. The tumours are highlighted using the yellow dotted lines. The left and right figures show a maximum depth of 5cm, while the middle figure shows a maximum depth of 4.5cm. 150

8.7 Cross section of one of the ovine brains. In this sample, two sites were injected with agar (dyed yellow) 150

8.8 The right example shows a real case of a brain tumour, highlighted in yellow, with a necrotic core, highlighted in red. On the left is water added to the agar to simulate a necrotic core. Yellow line outlines the agar, red line outlines the water. 151

8.9 Samples from the dataset with super imposed bounding box annotation. 155

B.1 Medical imaging modality deep learning research output trends. 209

B.2 Annulus sector geometric properties, created using `https://www.geogebra.org`. 212

B.3 Extracting the US plane. Left shows the original image. Middle shows a plot of the z-score, where in this circumstance there is only one spike as the background pixel is constant at 0. Right shows the output of the connected components process, where the US plane is the largest object and represented in white (the other colours represent other objects, belonging to the GUI). 213

B.4 (Right) An example of the calculation of the vertical axis of symmetry using the sliding window method. The blue line represents the intensities of vector S, the red line is the vertical line passing through point m and the green lines are the vertical lines through points min_l, min_r. (Left) The binarised US plane with the same lines superimposed. 215

B.5 Examples from the controlled experiment with the private data. The left images show the processing of the example images using the proposed method. The white plane is the output of the connected components process. The purple dots are the edge points of the mask, and the blue lines and orange crosses are the output of the RANSAC and the estimated radial boundaries. The middle and right images show the estimated and ground truth key points. 223

B.6 Example images from the public data experiment, one example from each public dataset used. The left image is the output of the method, the middle is the estimated key points and the right is the ground truth. All examples highlight the issue of cropping on the top of the US plane. The POCUS and DUKE examples show the issue of US plane corruption. Both CLUST examples suffer from slight rotation. 224

B.7 Comparison of the circularity of the object in the original convex US plane compared to the linearised version. Quantitatively this comparison is measured using circularity and Proscrutes metrics. 226

B.8 Sample from POCUS dataset [BWC+21a] showing reverberation. The linearisation process correctly projects the reverberation as vertical. 226

B.9 Radio-frequency data validation. Left is the captured convex image. Middle is the ground truth linear image. Right is the linearised image estimated with our proposed method. 228

Chapter 1

Introduction

1.1 Motivations

To provide an accessible introduction, this section is kept concise and free of excessive technical detail. For a more in-depth background on the subject matter, refer to Chapter 2.

Although a rare form of cancer, brain tumours are one of the leading causes of cancer-related deaths among young adults. Their relative rarity, in combination with the complexity of the treatment and the general poor survival rates, has contributed to stunted developments in translational research.

In particular, for brain surgery, maximal safe resection is of the utmost importance - no excess and critical function tissue should be damaged. Typically, preoperative images, which are always captured for diagnosis, are used to guide the surgeon during resection. However, preoperative images suffer from reduced reliability during surgery, due to significant confounders such as brain shift. These factors emphasise that there is a serious need to develop methods to support neurosurgeons during surgery to accommodate the complex and dynamic surgical scene. Imaging and sensing that can be performed intraoperatively therefore present themselves as a key point of interest.

Developments in intraoperative imaging tools and techniques have generally seen progress,

but have been slow, with limited funding and research output. Magnetic resonance imaging (MRI) has historically been the modality of choice due to its diagnostic familiarity and superior image quality. However, intraoperative magnetic resonance imaging (iMRI) has shown limited promise due to its high cost, the need for dedicated equipment and specialised operating room setups, prolonged surgical durations, and associated anesthesiology complications. In contrast, while ultrasound (US) provides lower-resolution images and captures smaller spatial areas, intraoperative ultrasound (IOUS) offers a real-time, cost-effective alternative that integrates more easily into the surgical workflow. Given these advantages, this PhD focusses on the advancement of computer-assisted methods for IOUS, with the aim of improving its utility in brain tumour surgery and fostering greater interest in this field from both the clinical and technical communities. The secondary direction of this PhD is dictated by the aim of the Imperial–Technische Universität München (TUM) Joint Academy of Doctoral Studies (JADS) project, which is the building of a robotic system to automate the general IOUS process.

1.1.1 Healthcare Benefits of Robotic and Computer-Assisted Surgery

The ultimate goal of the contributions in this thesis, and research within this domain, is to support and achieve clinical translation that has a positive real-world impact. Unlike theoretical research, applied research that prioritises purely benchmark-gaming and research over practical outcomes risks offering little more than intellectual nothings. As such, before diving into the relevant literature and technical contributions of this thesis, it is important to highlight how the general scope of what is being aimed to achieve may lead to positive outcomes and application within the healthcare sector. For in-depth details of the clinical and technical background, please refer to Section 2.

Although it is difficult to quantify how "difficult" neurosurgery is, or whether it is any more or less so than other surgical specialisations, it is measurable that medical students perceive it that way [HDK11, BA19]. More importantly, whatever the cause, there is a global shortage of neurosurgeons [WTN+23], and theatre-attending neuroradiologists are rarer still [RCR24]. Therefore, it must be emphasised that this deficit in specialists will invariably increase individ-

ual workloads [INB+25], negatively affecting the probability of surgical error [SBB+10].

Intraoperative sensing and imaging is now a regularly recommended tool for brain tumour surgery to improve the outcome. For example, iMRI and IOUS are both recommended by the National Institute for Health and Care Excellence during glioma resection [Nat21]. However, the standardisation of neurosurgeon training for IOUS is limited [DLGS+22] - which has impacted the wide adoption of it. Furthermore, performing IOUS is an additional cognitive burden for surgeons [BJT+21].

Considering these factors, and as the global population is increasing [GAD21], along with the increase in the incidence rates of brain cancer [II23], it is vital to explore all avenues that can alleviate the burden on neuro-healthcare professionals.

Trends show that robotic devices and robotic assisted surgeries have become more common in hospitals [SCD20]. As such, studies have been performed within real surgery, showing that robotic assistance has provided positive benefits to surgeons. For example, a study measured that during the phases of high cognitive load and time pressure of laparoscopic surgery, robotic assistance has been associated with improved surgical performance [SMR+18]. More generally, NHS England forecasts that robotic-assisted surgery will become far more common, reducing the physical demands of complex procedures, increasing the number of complex surgeries per day, and improving patient outcomes [NHS25].

Furthermore, early investments and speculation have highlighted the potential benefit of robotic assisted surgery for low- and middle-income countries (LMIC). Due to the limited access to healthcare specialists in certain regions [HLK+15], telesurgery can provide an invaluable solution to this problem by bypassing cost and logistical challenges and having access to specialists through an online connection [Tam21] [MAA+22]. In addition, due to the disproportionate burden of infectious diseases in LMIC [GB], limiting the risk to surgical personnel and patients is vital. Robotic surgery can be a solution to control and minimise disease spread [MRMK20] through the ability to increase the distance between the patient and the staff (also intra-staff) or completely separate the patient; potentially, also reducing the number of staff required altogether.

1.2 Thesis Structure And Summary

Each chapter is based on an individual project and publication. The chapters are grouped into three categories: Pathological Tissue Delineation (Chapters 3, 4, 5), challenges with interpreting brain tumours in US, particularly concerning tumour margin uncertainty; Probe Manipulation (Chapter 6), evaluating US probe interaction with brain tissue; Resource Leveraging (Chapters 7, 8), key areas where data and simulation resources are limited, which has affected translation efforts.

Figure 1.1 situates this thesis within the wider research space of robotic IOUS for brain tumour surgery and highlights the pertinent technical and clinical domains. Part I establishes the clinical need by addressing the key task of tissue characterisation: interobserver-variability study, a pragmatic sparse-annotation method, and robust fine-tuning for limited data. Part II demonstrates the feasibility of robotic IOUS for brain surgery using contact-aware ultrasound feedback to manipulate the probe. Part III mitigates resource limitations via two contributions: a realistic, low-cost ex vivo brain phantom and a data-efficient 3D reconstruction framework. Although complete end-to-end integration lies beyond the scope of this thesis, these contributions constitute foundational steps towards a complete system and an early exploration of fully automated IOUS for brain tumour surgery. Note: the narrative order is I→II→III for clarity and emphasis on the more fundamental areas, while the logical development and system flow runs inverse III (resources)→II (guidance)→I (analysis). This workflow has been highlighted in Figure 1.2.

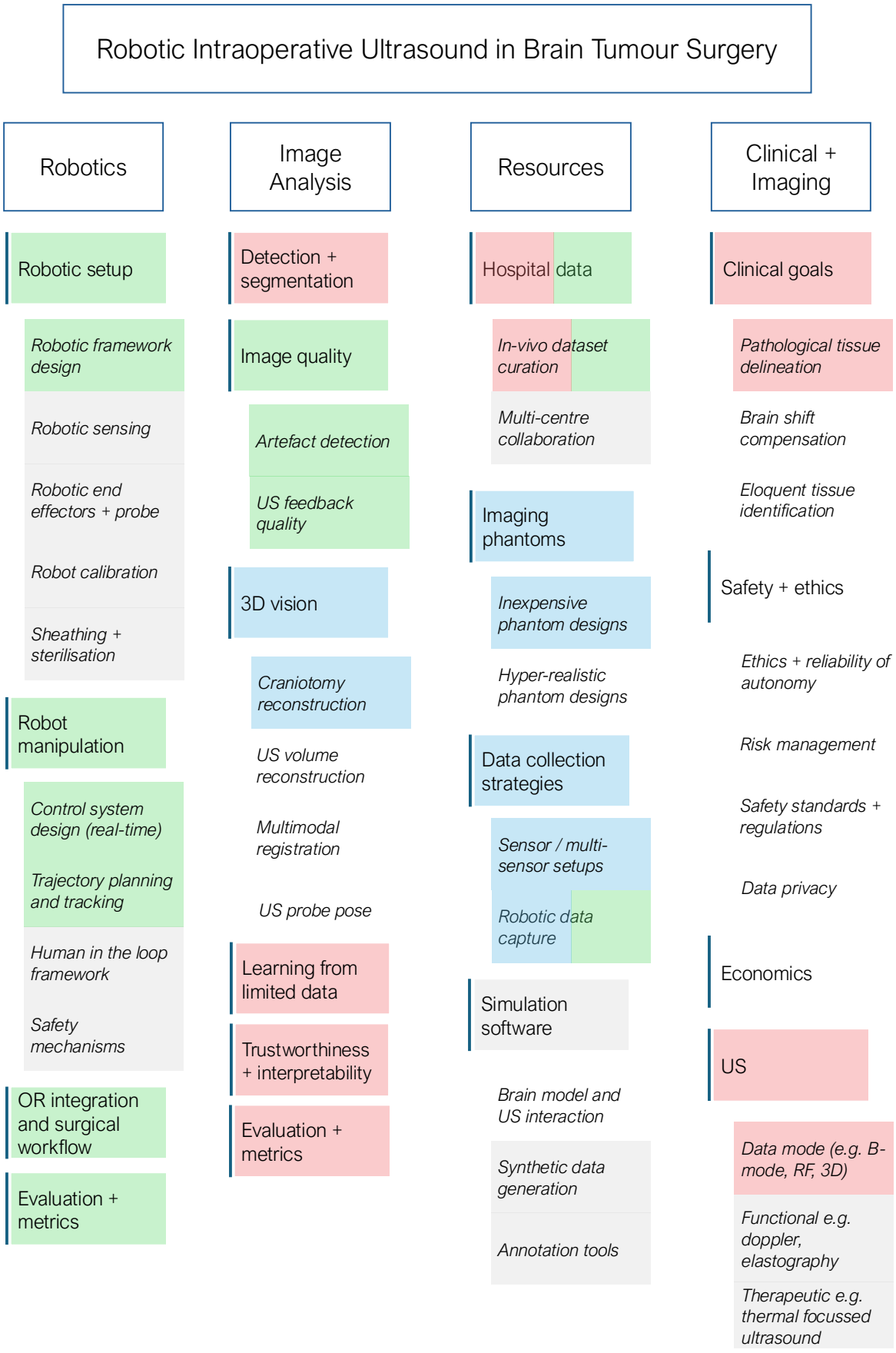


Figure 1.1: The research space for robotic intraoperative ultrasound in brain surgery. Subdivided into four key areas (Robotics, Image Analysis, Resources, Clinical goals + Imaging) and their respective subtopics . The colours of the boxes indicate: red = Part 1; green = Part 2; blue = Part 3; grey = topics addressed during the PhD but not included in the thesis.

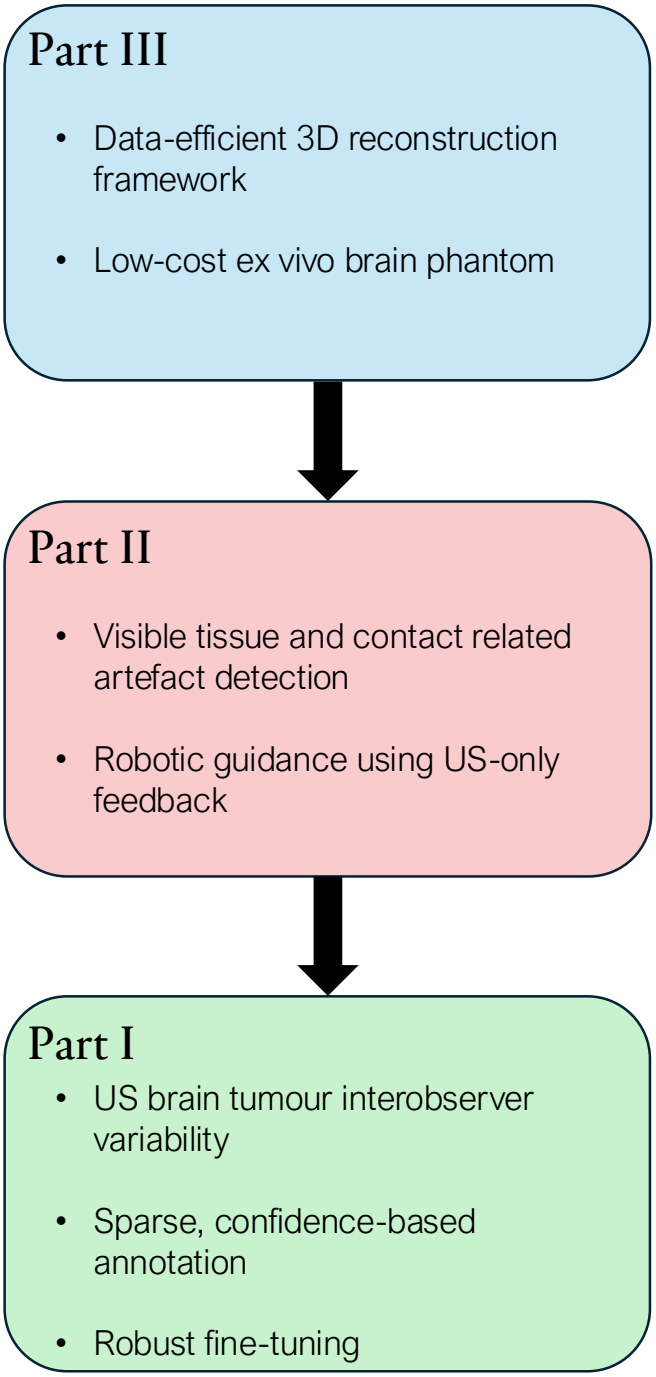


Figure 1.2: Thesis workflow and system logic for robotic IOUS in brain tumour surgery. The logic proceeds from methods that enable experimentation (Part III) to ultrasound-driven probe–tissue coupling (Part II), and finally to pathological tissue analysis within the acquired images (Part I). The box colours are the same as in Figure 1.1.

Chapter 2: Background. A broad background is provided on the clinical and technical scope of intraoperative ultrasound for brain tumour surgery. The clinical section describes the basics of brain tumours, the surgical process, and imaging techniques (not limited to ultrasound) used before and during surgery. The technical section describes the main areas of current and relevant research on this topic.

Chapter 3: Interobserver Variance Of Tumour Annotation The baseplate method for annotation is segmentation, be it multiclass or binary. Brain tumours are infiltrative with poorly defined margins, raising the question of whether or not it is possible to segment brain tumours in US images. In this chapter a study is presented that focusses on measuring the significance of the delineation uncertainty, due to the lack of resolution of ultrasound and the lack of formal and standardised clinical training.

Chapter 4: Robust Fine-Tuning For Tumour Detection. It can be assumed that within the near future most medical challenges will still lack appropriate volumes of available data with accurate and reliable annotations for standard supervised deep learning training. This chapter explores the benefit of fine-tuning using a newly proposed constrained gradient descent method that leverages anchoring on pre-trained parameters.

Chapter 5: A Sparse Annotation Method Based On Confidence Grading The binary classification of brain tissue goes against what is known about the biological properties of brain tumours. The is specifically the case along the tumour margins where cancers behave infiltratively. This chapter contains a proposed method for sparse annotation of tumours, based on spatial-wise confidence grading.

Chapter 6: A Method For Evaluating Ultrasound Probe-Tissue Contact During Brain Surgery. During US scanning, a professional user will reposition and reorient a probe using purely image feedback. A method is developed to replicate this process, specifically by identifying the visible tissue which can subsequently be used to infer contact related issues with the probe-tissue coupling, providing usable information for reorientation guidance.

Chapter 7: Learning Structured Light Projections To Improve Stereo Disparity

Learning From Limited Data. The extraction of high resolution 3D information of a surgical scene is particularly beneficial for ensuring appropriate contact between surgical devices and tools and the tissue surface. In particular, for defining US probe trajectories, whether it be towards the brain surface or along the surface. Available surgical data is limited, ethics are hard to acquire, and the capture of precise and reliable in vivo ground truth is difficult to achieve. A training framework is proposed that uses joint learning of structured light, to supplement limited-depth training datasets for improved learning from limited data.

Chapter 8: Inexpensive Brain Model With Tumour Simulation. Commercially available and homemade methods to create brain phantoms with accurate anatomical imaging in US are limited. This chapter explores a method for using inexpensive ex vivo brain tissue and simulating tumours using injected agar.

Chapter 9: Conclusion. A conclusion of the thesis is provided with comments on the limitations of the work and possible future directions.

1.3 Publications

1.3.1 Thesis Specific Publications

These publications are directly relevant to the subject of this thesis. All publications are open-access articles distributed under the terms of the Creative Commons CC BY licence. This permits unrestricted use, distribution, and reproduction in any medium, provided that they have been properly cited. All publications can also be found on Arχiv and medχiv.

- **Chapter 3** -

  Weld, A.*, Dixon, L.*, Anichini, G.*, Patel, N., Nimer, A., Dyck, M., ONeill, K., Lim, A., Giannarou, S., & Camp, S. (2023). Challenges with segmenting intraoperative ultrasound for brain tumours. Acta Neurochirurgica, 166.

- **Chapter 4** -

  Weld, A., Dixon, L., Camp, S., & Giannarou, S. (2025). A Study Of Robust Fine Tuning Of DETR Models From Small Data For Tumour Detection In 2D Ultrasound. (Submission - TBD)

- **Chapter 5** -

  Weld, A., Dixon, L., Camp, S., & Giannarou, S. (2025). Confidence-Based Annotation Of Brain Tumours In Ultrasound. Arxiv. (Submission - TBD)

- **Chapter 6** -

  Weld, A., Dixon, L., Anichini, G., Dyck, M., Ranne, A., Camp, S.J., & Giannarou, S. (2023). Identifying Visible Tissue in Intraoperative Ultrasound Images during Brain Surgery: A Method and Application. International Journal of Computer Assisted Radiology and Surgery (IJCARS).

- **Chapter 7** -

  Weld, A., Cartucho, J., Xu, C., Davids, J., & Giannarou, S. (2022). Regularising disparity estimation via multi task learning with structured light reconstruction. Computer

Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, 11, 1206 - 1214.

- **Chapter 8** -

  Weld, A., Dixon, L., Anichini, G., Faoro, G., Menciassi, A., Camp, S., & Giannarou, S. (2024). A Method For Mimicking Tumour Tissue In Brain Ex-Vivo Ultrasound For Research Application And Clinical Training. Acta Neurochirurgica, 167.

1.3.2 Other Publications

Publications that are not specific to this thesis are listed below. It is worth noting that a portion of the work conducted in collaboration with colleagues in Germany has been excluded from the main body of this thesis to avoid duplication or overlap of contributions. Additional output from this collaboration, which remains unpublished, is expected in the near future.

- **Ultrasound Segmentation Using A 2D UNet With Bayesian Volumetric Support**

  Weld, A.*, Agrawal, A.*, & Giannarou, S. (2022). CuRIOUS/KiPA/MELA@MICCAI.

- **Deep Regression With Spatial-Frequency Feature Coupling And Image Synthesis For Robot-Assisted Endomicroscopy**

  Xu, C., Roddan, A., Davids, J., **Weld, A.**, Xu, H., & Giannarou, S. (2022). International Conference on Medical Image Computing and Computer-Assisted Intervention.

- **Toward Safe And Collaborative Robotic Ultrasound Tissue Scanning In Neurosurgery**

  Dyck, M.*, **Weld, A.***, Klodmann, J., Kirst, A., Dixon, L., Anichini, G., Camp, S., Albu-Schaffer, A., & Giannarou, S. (2023). IEEE Transactions on Medical Robotics and Bionics, 6, 64-67.

- **SurgT challenge: Benchmark Of Soft-Tissue Trackers For Robotic Surgery**

  Cartucho, J., **Weld, A.**, Tukra, S., et al. (2023). Medical image analysis, 91, 102985.

- **Probe-Based Raman Spectroscopy In The Presence Of Visible Light For Brain Tumour Analysis**

  Davids, J., **Weld, A.**, Hill, C., Thompson, A., Giannarou, S., Ashrafian, H., & Darzi, A. (2024). Journal of Neurology, Neurosurgery & Psychiatry.

- **NeuroMetaverses For Neurology, Neurosurgery And Psychiatry**

  Esposito, L.*, Davids, J.*, Miriyala, R., Berstecher, R., Jose, I., **Weld, A.**, Samarasekara, J., Adams, H., & Ashrafian, H. (2024). Journal of Neurology, Neurosurgery & Psychiatry.

- **2D Ultrasound Thermometry Based On Echo Nonlinearities And Deep Learning Denoising**

  Faoro, G., **Weld, A.**, Giannarou, S., & Menciassi, A. (2024). (Submitted to IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control).

- **Intraoperative superb microvascular ultrasound imaging in glioma: novel quantitative analysis correlates with tumour grade**

  Dixon, L., **Weld, A.**, Bhagawati, D., Patel, N., Giannarou, S., Grech-Sollars, M., Lim, A., & Camp, S. (2025). Acta Neurochirurgica vol. 167.

- **Multimodal Imaging Platform For Enhanced Tumor Resection In Neurosurgery: Integrating Hyperspectral And pCLE Technologies**

  Roddan, A., Czempiel, T., Xu, C., Xu, H., **Weld, A.**, Chalau, V., Anichini, G., Elson, D., & Giannarou, S. (2024). International Journal of Computer Assisted Radiology and Surgery (IJCARS) 20, 1087–1096.

- **Deep Learning-Based Glioma Segmentation on 2D Intraoperative Ultrasound Images: A Multicenter Study from the Brain Tumor Intraoperative Ultrasound Database (BraTioUS)**

  Cepeda, S., Esteban-Sinovas, O., Singh, O., Moiyadi, A., Dixon, L., **Weld, A.**, et al. (2025). Cancers, 17, 315.

- **SurgRIPE challenge: Benchmark of Surgical Robot Instrument Pose Estimation**

Xu, H., **Weld, A.**, Xu, C., et al. (2025). Arxiv. (Accepted at Medical Image Analysis)

- **Standardisation of Convex Ultrasound Data Through Geometric Analysis and Augmentation** -

  Weld, A., Faoro, G., Dixon, L., Camp, S., Menciassi, A., & Giannarou, S. (2025). Arxiv. (Submission - TBD)

1.3.3 Presentations and Posters

- **Weld, A.***, Dyck, M.*, Klodmann, J., Anichini, G., Dixon, L., Camp, S.J., Albu-Schäffer, A., & Giannarou, S. (2022). Collaborative Robotic Ultrasound Tissue Scanning for Surgical Resection Guidance in Neurosurgery. Proceedings of The 14th Hamlyn Symposium on Medical Robotics.

  Contribution - Co-first author.

- Dyck, M.*, **Weld, A.***, Klodmann, J., Kirst, A., Anichini, G., Dixon, L., Camp, S.J., Giannarou, S., & Albu-Schäffer, A. (2023). Automated Robotic Intraoperative Ultrasound for Brain Surgery. ICRA-RAMI workshop.

  Contribution - Co-first author.

- **Weld, A.**, Davids, J., Ashrafian, H., & Giannarou, S. (2023). AI-guided Ultrasonographic Assessment for Intraoperative Glioma Resection. Brain Tumour North West Conference.

  Contribution - First author.

- **Weld, A.**, Dixon, L., Anichini, G., Giannarou, S., & Camp, S. (2024). Intraoperative Ultrasound Brain Tumour Margin Uncertainty: Interobserver Variation And A Method For Annotation. Cancer Research UK Brain Tumour Conference.

  Contribution - First author.

- **Weld, A.**, Dyck, M., Dixon, L., Anichini, G., Camp, S., Albu-Schäffer, A., & Giannarou, S. (2024). Robot-assisted Ultrasound Brain Tissue Scanning for Surgical Resection Guid-

ance. Cancer Research UK Brain Tumour Conference.

Contribution - First author.

1.4 Coronavirus Disease 2019 Impact Statement

Between March 2020 and December 2021, the UK government imposed lockdowns and measures due to the 2019 coronavirus pandemic. The social and economic impact of this period was significant and continued long afterward. This PhD began early in January 2021. At least the first year of this PhD was heavily affected by the global pandemic. Office work was challenging and many colleagues did not return to the office until long into 2022. Collaboration with colleagues working within the NHS and abroad in Germany was difficult. In particular, the NHS had little time and money for research for the duration of the PhD. The principal objective of this PhD is clinical translation research and therefore the pandemic greatly affected the shape and direction of the work in this thesis. In particular, all 3D vision work was performed during the pandemic. Face masks required and all.

1.5 Descriptions, Conventions and Short Handings

To help with readability, equations are redefined within each chapter to maintain smooth notation and contextual relevance. This approach keeps the notation self-contained within the chapters and ensures closer proximity for the reader to refer to. Abbreviations, such as ultrasound (US), are also redefined within each chapter for the same reason.

Chapter 2

Literature Review

This review of the background literature presents the clinical challenges and key technical developments in the space of computer-assisted intraoperative ultrasound for brain tumour surgery. Since each contribution chapter corresponds to an individual research project and includes its respective relevant literature review, this chapter minimises redundancy by discussing the broader subject and highlighting areas of interest - both covered and not covered later in this thesis - that I perceive as being relevant, important, and potential avenues of future work.

2.1 Clinical Background

2.1.1 Brain Tumours

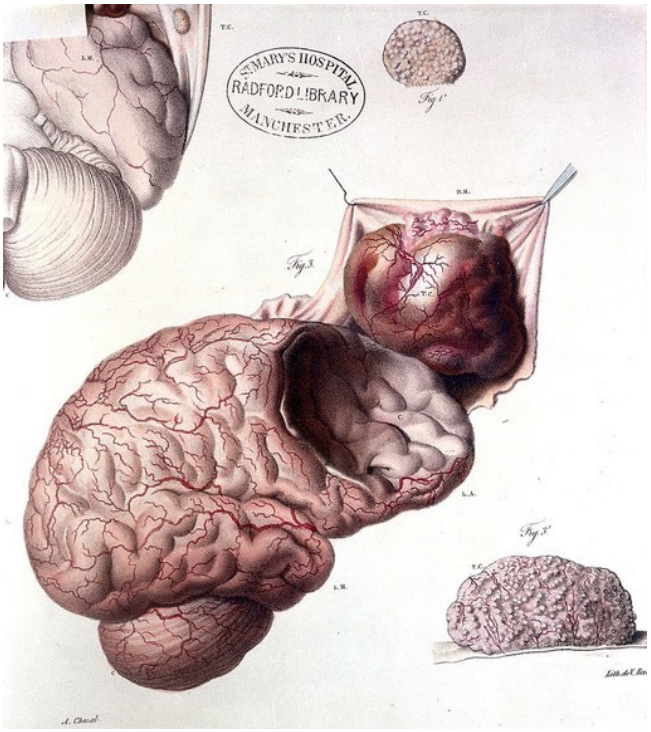


Figure 2.1: Brain tumour illustration - Wellcome Collection, United Kingdom - CC BY. https://www.europeana.eu/item/9200579/use8h2y8

Brain tumours are intracranial neoplasms, abnormal growths of cells that can form in or around brain tissue [DeA01]. An illustrative example is shown in Figure. 2.3 They can be benign (noncancerous) or malignant (cancerous). They may originate from the brain (primary) or have spread from other cancers (secondary / metastatic), such as the central nervous system (CNS). The international standardisation of CNS tumour classification has been dictated by the WHO. Where the classification is determined by cell of origin or anatomical origin, resulting in the identification of 120 types of brain tumours. The current standard was decreed 2021 in the 5th edition of the "WHO classification of CNS tumours" [LPW+21].

According to Cancer Research UK, in the UK, measured between 2017 and 2019, approximately 12,700 new cases were diagnosed every year [cru]. This continues the trend observed over the last decade of an increase of 10% of the incidence rate. Where, the rate of new cases was highest among those between 85 and 89 years of age. However, the distribution of tumour type varies

significantly across age brackets [MPL10]. Between 2017 and 2019, brain tumours represented 3% of cancer deaths in the UK. During the three years, 5456 deaths were associated with brain tumours, which is approximately equivalent to 14% of the total number of new cases.

Although brain tumours are less common than other cancers and rank lower in overall cancer-related mortality in all age groups, they are the leading cause of cancer-related deaths among children and young adults under 40 years of age since 2002, according to Brain Tumour Research [Bra] [RS23]. Despite this, only 1% of national cancer research funding is allocated to brain tumour research.

The diagnosis and treatment planning of brain tumours is strongly determined using magnetic resonance imaging (MRI), which has replaced invasive historical methods such as lumbar puncture [LT93]. MRI provides noninvasive high-resolution imaging with good soft tissue contrast. More significantly, the pulse setting can be varied in MRI to achieve different imaging sequences, which can expose different perspectives of the tumour embedding and interaction with surrounding normal tissue [MRS+23].

An example of commonly used MRI sequences is the set present in the BRATS dataset [MJB+15], a seminal benchmark dataset for brain tumour MRI image segmentation. These sequences included are: T1-weighted (T1), T2-weighted (T2), contrast-enhanced T1-weighted (T1CE) and T2 Flair (FLAIR). Images from the dataset are shown in Figure. 2.2. Additional common sequences that are used, but not part of BRATS, include diffusion-weighted and perfusion-weighted.

Furthermore, due to the complexity of pathological invasion into the surrounding normal tissue adjacent to the tumour, the limits of what should and can be resected are difficult to determine. For example, approximately 7% of the gliomas will have invaded eloquent regions, critical functional tissue, such as speech [HS03] [SOB+96]. In such cases, functional MRI, a sequence less commonly used, becomes crucial to providing insight by mapping the interaction of the tumour with the eloquent cortical regions [LSC+23].

In addition to MRI, Computed Tomography (CT) and positron emission tomography (PET)

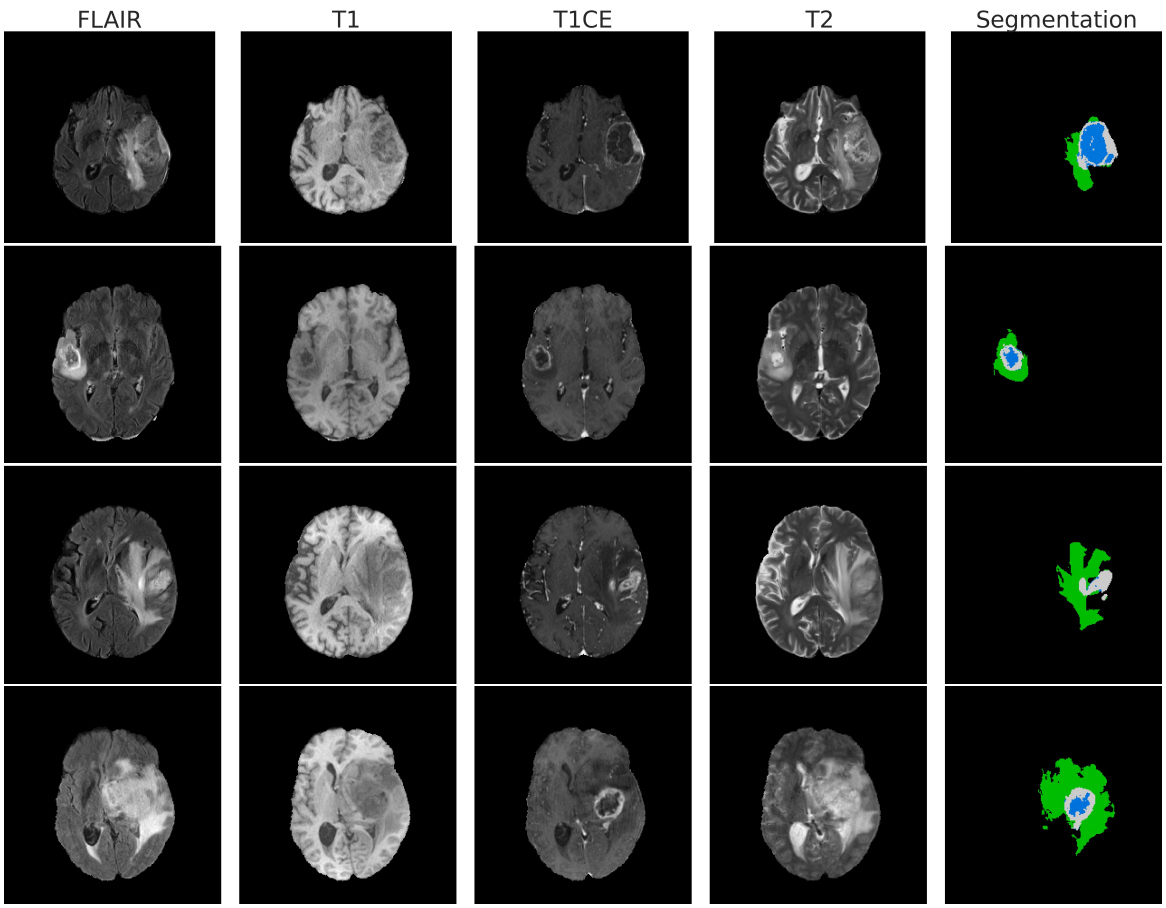


Figure 2.2: Example images from the BRATS dataset showing different MRI sequences. Segmentation legend: blue is necrotic and non-enhancing tumor core, green is peritumoral edema, silver is Gadolinium enhancing tumour.

images are also used to provide complementary information for further investigation of the tumour and also provide circumstantial information for which MRI might not be suitable. CT imaging is particularly useful for detecting bone damage, tumour calcifications, and haemorrhages, which can be secondary to tumours or confound their diagnosis. PET imaging provides advantages by capturing metabolic information, including tumour extent, differentiation of treatment-related changes from recurrent tumour, and does not rely on blood-brain barrier breakdown, unlike contrast-enhanced MRI [LCS+09].

2.1.2 Brain Tumour Surgery

Treatment of tumours will be highly dependent on the centre, patient, and tumour specifics [PL16]. However, treatment usually involves a combination of radiation therapy, chemother-

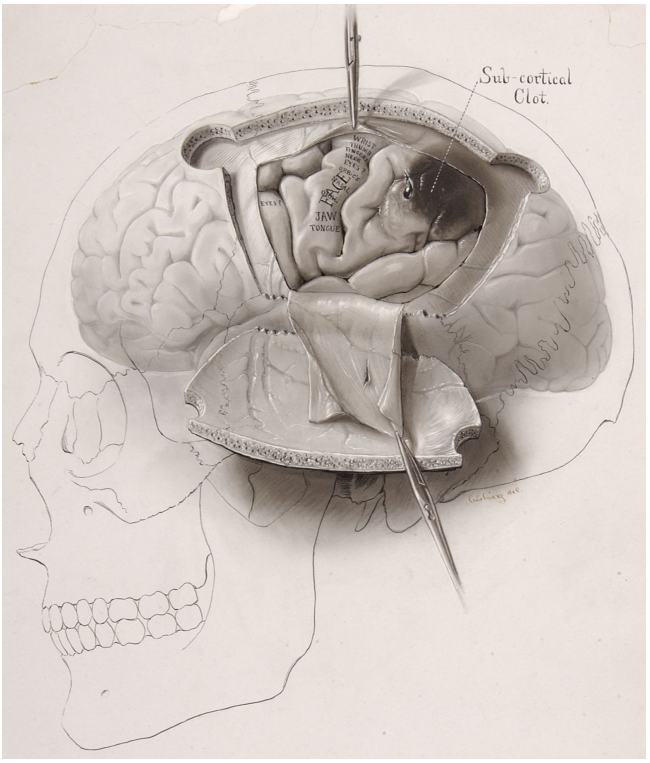


Figure 2.3: Illustration of a craniotomy process - "Drawing of the brain showing a subcortical clot." from Harvey Cushing Photograph Collection, Medical Historical Library, Cushing/Whitney Medical Library, Yale University Library Digital Collections, Image ID: 32325691, https://collections.library.yale.edu/catalog/15951324.

apy, and surgery. Where surgery is usually the first and most common treatment approach. Specifically, the purpose of surgery is the maximal safe resection of pathological tissue.

Glioblastomas are one of the most insidious brain tumours due to their aggressive nature and their complexity in biological properties and treatment. Studies of the extent of resection, specifically comparisons of gross total resection (GTR) with subtotal resection (STR) - where GTR is the removal of all visible tumour (identified on MRI), whereas STR leaves residual pathological tissue due to interaction with eloquent cortex, in order to minimise the risk of iatrogenic complications - have shown that GTR probably improves survival probabilities, particularly for 1 and 2 years by a significant factor [BBL+16]. However, it is important to acknowledge that these studies are substantially limited by poor control of patient and tumour-specific factors, and that the studies are retrospective. Furthermore, confounders such as tumour size, interaction with eloquent structures, and postoperative therapies contribute significantly to determining whether GTR or STR is pursued and therefore also mortality probabilities. This

is highlighted in the cited systematic review and meta-analysis. Regardless of the situation, what is clearly highlighted through this is that surgery is fundamental to treatment, and the complexity of the surgery is a major factor in the likelihood of patient mortality.

The discussion of the surgical process is beyond the scope of this thesis. As such, I will reduce this process to merely highlighting the mechanical aspects. That is, scalp incision, removal of the bone flap and cutting of the dura material [CVB08]. This process is shown in Figure. 2.3. At this point of the procedure, confounding factors begin to appear that complicate the reliability of presurgical information such as preoperative imaging. Which is what the main context of this thesis is concerned with.

There are a multitude of variables that occur post-craniotomy, with the main contributing factor being intracranial decompression, as the brain is normally contained within a protected and closed environment. The most noticeable factor is brain shift, which is the issue of brain deformation during the surgical process. For example, brain tissue and also tumour tissue are known to increase in volume after craniotomy [KDBF+24]. In addition, the brain itself may have a different intracranial orientation due to simple factors such as the position of the patient and the effect of gravity [GKOH+21]. Furthermore, during cutting, the tissue gradually undergoes an increase in deformation, which correspondingly reduces the preoperative MRI's representation of the current surgical scene [RC15].

2.1.3 Intraopeartive Imaging

The natural solution to the reduced reliability of preoperative images is to reimage during surgery, this process is called intraoperative imaging. The most explored imaging modality is intraoperative MRI (iMRI) [MGB10]. This is due to the familiarity with the modality during the diagnostic stage, as well as to the high-quality visualisation of the tumour tissue and the entire 3D brain structure.

However, logistical concerns and preparation requirements make iMRI infeasible as an on-the-fly tool. A major limiting factor is the added complexity of anaesthesia management. From

this study [GB18], it is implied that iMRI can only be performed if the number of scans and total duration are pre-determined preoperatively. This encumbersome scheduling requirement prevents its flexible use during surgery. Furthermore, a study [BJF+09] reported that setting up a dedicated iMRI system (POLESTAR™) and operating room can take anywhere from 30 minutes to 3 hours, with image acquisition times ranging between 8 seconds and 13 minutes. Beyond time constraints, the financial burden is significant: another study [MCE+11] calculated the initial investment and 5-year running cost of their system to be $3.8 million.

Additional imaging and optics alternatives include hyperspectral and multispectral imaging [ALH+24], Raman spectroscopy [ZqYL+23], computed tomography [ACH+20], and confocal laser microscopy [RPV+21]. However, none have achieved any standardisation and are all in the infancy of investigation for practical application.

2.1.4 Ultrasound

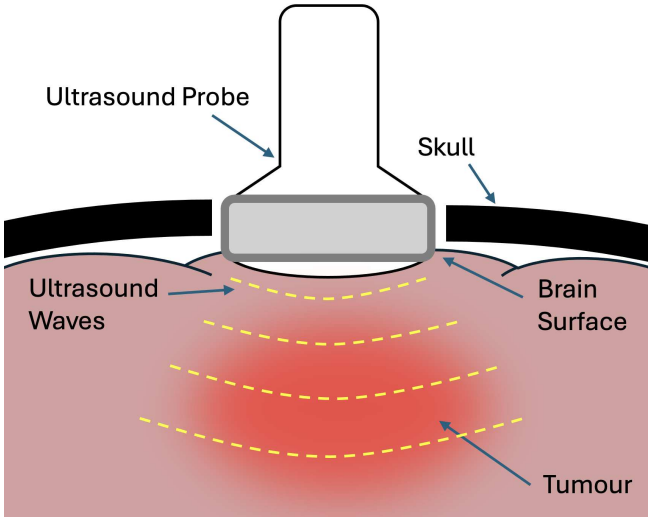


Figure 2.4: Simplified diagram of the US scan within a craniotomy.

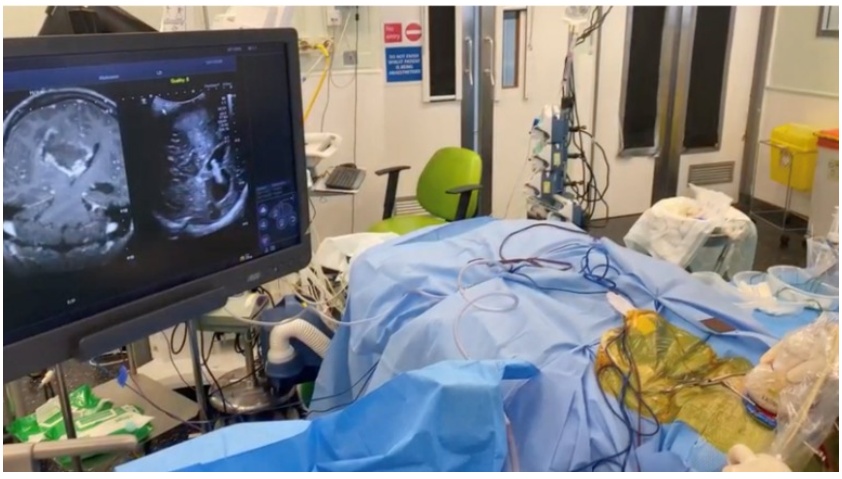
Arguably, the most promising intraoperative tool is ultrasound (US). Where by US, I mean medical US devices that emit ultrasonic signals in the diagnostic range of ≈ 2–20 MHz, from a piezoelectric matrix (transducer) within a handheld probe. The transducer converts electrical energy into high-frequency sound waves that propagate through tissue and reflect back to the transducer, at acoustic boundaries. These reflections, known as echoes, will vary in amplitude



(a) Convex probe ultrasound scan through a small craniotomy window.



(b) Scan inside resection cavity.



(c) Ultrasound manual registration with preoperative MRI.

Figure 2.5: Intraopertative ultrasound scan captured at Charing Cross Hospital.

and density depending on the impedance properties of the boundaries (echogenicity) and possible artefacts. The echoes are received as radio frequency (RF) signals, which, for standard clinical devices, are immediately processed and converted into a B (brightness) mode grayscale image; the RF signals are usually inaccessible outside of research devices. The depth of the image corresponds to the depth of propagation of the signal, while the brightness varies with the echogenicity of the tissue, for example, hyperechoic (bright) for bone and hypoechoic (dark) for the abscess core [PP24].

There is a wide array of US probe types available. For neurosurgery, the most relevant probes are linear, convex, phased array, and hockey stick. Where the convex and phased array probes can also capture 3D US volumes. However, since 3D probes offer images of much lower quality than 2D [AEP+22], due to the topic of this thesis, the work is focussed on 2D. This was verified with our clinical partners and exploration with different machines and probe types.

Intraoperative ultrasound (IOUS) is the application of US post craniotomy, before, during, and after resection. A simple diagram outlining the US process, within a craniotomy, is shown in Figure 2.4. Pictures captured during brain surgery at Charing Cross Hospital are shown in Figure 2.5. The benefit of IOUS, particularly over iMRI, and why this tool is the centre piece of this thesis, is that IOUS can be used on the fly for acquisition of real-time information with minimal extension of the surgical time taken, progressively improving image quality, and enormously cheaper (orders of magnitude, low hundreds of thousands compared to millions) initial investment cost and running cost. Although iMRI does provide better image resolution and tissue differentiation - particularly at the end of resection - the pragmatics involved make it significantly more difficult to establish it as a widely standardised tool across centres and countries, for neurosurgical oncology [MSR+22].

Notable factors in the slow adoption of IOUS include the perceived difficulty in mastering (e.g. image interpretation and visuotactile skills), standardisation, and the large variability of US machines and probes [DLGS+22]. In particular, the lack of standardisation and formal training in the use of IOUS has meant that studies evaluating the use of IOUS may have been conducted by those without sufficient training and experience with the tool.

2.2 Technical Background - Computer Vision Algorithms For Limited Data Applications

Automated analysis of digital images is a mature field with numerous promising applications in clinical practice. Radiological, sonographic, and optical imaging tools are becoming increasingly important and widely adopted in surgical settings. These trends have motivated considerable interest in translating computer vision algorithms into operating theatres; across the medical, academic, and commercial sectors.

The application of machine learning algorithms for surgical tasks has now begun to seriously demonstrate translation viability [VHOT24]. Mostly, these algorithms are based on deep learning. One of the first breakthroughs was a 2018 collaboration effort between DeepMind and Moorfields Eye Hospital, for the diagnosis and referral decision for retinal disease [FLRP+18]. In this work, they demonstrated that standard segmentation and classification neural networks could be trained to match or exceed the performance of experienced clinicians. In addition, another contribution was the demonstration of the ability to retrain segmentation neural networks, without prohibitive data requirements, to adapt to another optical coherence tomography device. Although this was a significant breakthrough in demonstrating the potential of neural networks, it should not be obfuscated that the key to deep learning success is access to large datasets. In the case of this work, the original training was performed on 14,884 images and it should be noted that the distribution shift between the two devices was relatively minimal. Furthermore, this study was industry-financed, the hospital is world-leading, the recording of retinal images is naturally part of clinical practice, and the diseases are relatively common. Some large medical image datasets are available, but when considering the total set of clinical tasks involving imaging, it must be remembered that there is fundamentally a scarcity of available data. Efforts are underway to improve this situation; however, these hurdles present an uphill battle, including the difficulty of acquiring ethics, the limitation of hospital resources, the limited incentives to share data, and the rarity of the conditions.

A brief historical outline of the key developments in machine learning is presented in Appendix

Chapter 9.2.3. I believe that knowing this historical context is critical to understanding the current state of research; notably, how we arrived here, which should provide a reference point to current data limitation problems.

The comprehensive exposition of the fundamentals of linear algebra, information theory, computer vision, and machine learning is purposefully omitted from this thesis, as these topics have been covered ad nauseam. For a comprehensive background on these subjects, the following seminal textbooks are recommended: Sheldon Axler's "Linear Algebra Done Right" [Axl95], Christopher Bishop's "Pattern Recognition and Machine Learning" [Bis07], Kevin Murphy's "Machine learning: a probabilistic perspective" [Mur12], Richard Szeliski's "Computer Vision: Algorithms and Applications" [Sze11], Richard Hartley's "Multiple View Geometry in Computer Vision" [HZ04], Ian Goodfellow's "Deep Learning" [GBC16].

2.2.1 Non-learning Examples in Surgical Application

When data is limited, or when explicit task and environment modelling is feasible, non-learning algorithms offer a strong alternative. Using natural information such as known physics or geometry, these methods may inherently offer superior interpretability, better behaviour under distribution shift, provable convergence (under specific assumptions), analytically derived uncertainty, and robustness compared to deep learning methods. These strengths make non-learning solutions advantageous for many surgical tasks, including surgical robotic guidance and diagnostic support.

For reference material, a brief list of examples is provided that were read during the PhD; for this and the deep learning section below, non-US examples will be provided to diversify the review of the literature. For 3D Vision: using Simultaneous Localization and Mapping (SLAM) for endoscopy tracking and mapping [MCH+16]; combining SLAM with shape-from-template and non-rigid structure-from-motion to perform deformable tracking and mapping of soft tissue surfaces in surgical videos [LPBM19]; point tracking of soft tissue features using salient feature extraction in conjunction with an Extended Kalman Filter [GSY13]; surgical robotic tool tracking and pose estimation using pattern markers [CWH+21]; surgical scene depth estimation

for robotic tissue scanning probe guidance using stereo vision depth estimation via a Bayesian stereo matching algorithm [ZCG20] [GRU10]. For Tissue Characterisation: polyp segmentation in endoscopy using principal component pursuit and active contours [SIBK18]; brain tumour segmentation using fuzzy integrated active contours [JRP19]. For Image Quality: uncertainty quantification in fast, undersampled MRI using Bayesian hypothesis testing [RPW18]; joint reconstruction and (structural) segmentation of fast, undersampled MRI using a Bregman iteration framework [CBE+18]; physics informed random walk for US confidence analysis, to guide robotic systems [KWKN12a]. For Registration: 2D IO camera with 3D preoperative MRI brain surface registration through elastic deformation, optimised with gradient-descent energy minimisation [BCH+14]; The registration of iMRI and preoperative MRI images for estimation of brain shift and assessment of residual and tumour margins, via non-rigid registration [CDT+05].

2.2.2 Deep Learning With Limited Data

Much work has been done to explore and quantify the problem of data limitation for deep learning training. The first question, particularly for safety-critical areas like medical imaging, is how much training data is sufficient [CLS+15]. While a logical starting point, in practice, answering this is inherently circular - to know the answer, you need to have the data to see how the performance will change; a catch-22. This has been highlighted in such papers as [DHM+20] which explore the idea of underspecification. That is, held-out data usually follow the same distribution as the training data and therefore bias judgement and evaluation on trained performance.

In [HNA+17] [BDK+21], empirical evaluation, across multiple data domains, shows that neural network performance, specifically generalisation error, decreases approximately as a power-law with the size of the training dataset.

It is clear, that work in this domain may indefinitely suffer from the limitations of available data. However, this should not prohibit attempts to do research and development with what is available.

Although most surgical imaging tasks lack large, organised annotated datasets, it is often feasible to collate small datasets manually or from various public sources. This can facilitate two paths of algorithm development: (i) calibration and evaluation of non-learning algorithms, and (ii) well-planned training of neural networks. Granted that fully supervised, standard neural network training given voluminous datasets will usually result in the best possible performance. The considered exploitation of small datasets within strategic training frameworks may still yield competitive performance. Various mechanisms to achieve this are listed below; this will include approaches that will range from improving learning given the limited data to methods that aim to synthetically generate data.

2.2.2.1 Data-Focussed Examples

Elemental training strategies involve the careful design of how data and labels are prepared and presented. Data-focussed methods can be divided into mainly three subcategories.

- Data processing - The deterministic curation of existing data to maximise information gain or representation of dataset labels. Fundamental examples may include batching and shaping of data. A well-known example of for the propagation of dataset labels is the "Stereo Correspondence and Reconstruction of Endoscopic Data Challenge" (SCARED) [AMW+21]. For this dataset, 45 dense depth map key frames were captured using porcine cadavers. However, through careful warping of the key frames onto intermediate frames using the da Vinci robot kinematics (which held the stereo endoscope), a dataset was created that contained 17,206 training and 5,637 test samples of rectified stereo pairs with depth ground truth. This dataset facilitated research into 3DV for surgical scenes.

- Data augmentation - On-the-fly modification of data during training to enrich the learning process and fill in and broaden the taught in-distribution. Common augmentation techniques include additive noise, smoothing, or compression; random cropping or zeroing; within-sample shuffling; geometric and colour-space transformations. Although not always physically realistic, these techniques have become commonplace in all deep learning and have demonstrated unquestionable benefit to improving the generalisation and

robustness of neural networks. In medical imaging, augmentation has shown improved efficacy when the physics of data acquisition or representation are obeyed. Examples of these augmentation techniques would include: spectral augmentation for improved calibrated learning of hyperspectral data [RCEG24]; generating motion artefacts in MRI through k-space filtering [SSV+20]; "copy-paste" tumours in kidney CT to improve segmentation training [YZL+21]. Although not tested within a training pipeline, the augmentation method highlighted in the Appendix Chapter B could also hypothetically be used to improve US learning.

- Data generation - The use of generative models to supplement datasets with semi-unique data samples. The history of generative models has been diverse, with many subtle variations in interpretations and applications. I refer to Appendix Chapter 9.2.3 for a better context of this. However, in modern deep learning paradigm, generative models fundamentally root to: Variational Auto Encoder (VAE) models [KW13], Generative Adversarial Networks (GANs) [GPAM+14], and diffusion models [HJA20a]. Where, overall, these models learn to approximate the probability distribution of given training data and can be used to generate new samples from the learned distribution. Generative models have shown unprecedented capability in recent years, and in many cases, synthetic images, audio, and videos have become difficult to distinguish from real [AMS22]. This qualitative impression of synthetic capabilities has become appealing within the medical community: "If i cannot see the difference, if it looks good enough, why should I not train a neural network with it?" Certainly, empirical evaluation in many different applications and published works has demonstrated the benefit of training using synthetic data [AKS+23] [KWA+23]. However, on a personal note, I believe it is clear that there are certain underlying issues which the research community has neglected in favour of advancing benchmark performances. These are both issues regarding the ethics of the use of synthetic data as a proxy to real data, and also the potential impact of performance under real distribution shift are yet to be fundamentally evaluated.

2.2.2.2 Transfer learning

Similarly to data augmentation, a common default strategy is to take a neural network with pre-trained weights, pre-trained on an alternative large dataset, or to perform this initial training step manually. enabling improved learning in target datasets / tasks.

In circumstances where alternative datasets are available, similar to the target dataset - for example, dataset A containing 100,000 images with classification ground truth of household pets, and dataset B containing 100 images exclusively of blue whales. Pre-training on the large dataset allows the neural network to learn parameters which can discriminate semantic features similar to what is expected in the target dataset. Therefore, continuing training from this primed position leverages a model with strong foundations, making increased performance and avoidance of overfitting more likely. Even if the two datasets contain fundamentally different information, there may still be unrealised common information, e.g., low-level features like edges or intensity patterns. Regardless of the distribution shift between the datasets, starting from a learnt position may inherently provide an advantage over starting from randomness.

An early investigation of the benefit of transfer learning in medical image applications was [SRG+16]. Here, the authors explore how transfer learning could be used with convolutional neural networks, applied to CT images, for lymph node detection and lung disease classification. The main outcome of this work was the consistent evidence of the benefit of using networks pre-trained on natural image data. A detailed investigation of the impact of transfer learning on neural networks, how the feature extraction process adapts, where the benefits could lie, and the limitations is provided in [RZKB19].

Subsequently, a large volume of research and commercial exploration of deep learning for medical imaging have relied on this training strategy as a matter of course. More nuanced are now the subcategories of transfer learning, which would include:

- Linear probing - A simple and fast to train technique that involves freezing a pre-trained backbone [EAAAM22] and appending a trainable linear output layer/head. This approach has demonstrated strong advantages, through keeping the feature extraction layers fixed,

thereby reducing the risk of overfitting. Using a foundation model [Com23] based on Contrastive Language–Image Pre-training (CLIP) [RKH+21] and pre-trained on biomedical vision–language data, the authors show that linear probing can achieve state-of-the-art performance for domain adaptation to limited-label classification tasks on benchmarks such as the RSNA pneumonia dataset [SWH+19] - with competitive performance on other datasets including PatchCamelyon [VLW+18] - even with as little as 1% and 10% labelled data [ZXU+23]

- Fine-tuning - Involves training a pre-trained network, either all layers or a subset, with small updates or small learning rates, to limit the overall change in the feature extraction process of the original learnt parameters. This enables the network to adapt to tasks with larger distribution shift than linear probing, while also providing a framework which can be designed to encourage the maintaining of performance on the original training distribution. Fine-tuning is a broad category in itself, therefore I recommend this detailed survey comparison of multiple fine-tuning strategies for medical image classification, involving multiple architectures and datasets, which also including comparisons against linear probing [DCH24].

- Parameter-Efficient Fine-Tuning (PEFT) - With the trend of successful training of neural networks with continuingly growing numbers of parameters [HGL+24]. Recent interest has moved to training strategies for updating large models through training of a subset of parameters. As generalisability of large models, notably large language models [VSP+17] [Ope22] and vision transformer (ViT) [DBK+20], have demonstrated increasing capabilities in understanding cross domains and modalities, this has struck interest in the medical community, as fundamentally, this replicates the required real-world decision-making protocols. A detailed study of PEFT algorithms for medical tasks was published in [DES+23]. In addition, it covers a large number of multiple architectures, tasks, and datasets. What was highlighted was the benefit of PEFT in low- to medium-scale datasets and the overall potential of this avenue of research and development which will be a hot focus in the medical image analysis community in the coming years.

2.2.2.3 Learning Without Ground Truth

When ground truth / labels are not available, feature representations of the task can be learnt by training the network to model the input itself. Two main subcategories can be defined within this field:

- Self-supervised - Learning that is supervised by input data, without additional annotated ground truth. A seminal example for this, which has shaped a lot of research for surgical guidance learning from limited data, is the left-right consistency depth estimator [GAB17] (although historically termed unsupervised); depth is learnt by finding the differentiable warping of the left image to the right and vice versa, without any disparity ground truth. This has been used as the foundational technique for numerous successful translations for endoscopic application through stereo disparity estimation [TG22]. A comprehensive evaluation of self-supervised methods for the classification of medical images was published in [HPJ+23]. A main take from this study was that self-supervised pre training of models for annotated limited tasks, in general, offered better model performance.

- Unsupervised - Self discovery of representations and structures within the data which can be used for discrimination, without any explicit target. A particularly important implementation of unsupervised learning is the use of autoencoders to measure reconstruction error and encoded representation shift during inference. Models such as the VAE are trained on healthy data to encode and then decode input data, where the decoded output should be identical to the input. This is done using data without the need for labels. Then, during inference, if the encoding and decoding process produces abnormal behaviour, this can be used to classify medical image anomalies such as tumours in MRI images [CSD+22]. For reconstruction, if the similarity measure against the input shows a significant difference, these data can be determined as anomalous. For the encoding, the VAE is trained to map data to normal distributions, so if the data is mapped to something anomalous to the normal distributions, this can be used to signal anomaly. A comprehensive survey of unsupervised learning using VAEs for medical imaging can be found in [USHE18].

2.2.2.4 Multi-Task Learning

Multi-task learning is the joint optimisation of a shared model for multiple related but different tasks, usually given the same data. The advantages for data-limited application are the regularisation impact during training and the maximisation of information gain from individual samples. Examples of joint optimisation frameworks include: histology classification and segmentation of different tissue regions [GVJ+22]; multi-modality foundation model for classification, detection, and segmentation [SNH+23]; capsule endoscopy pathology classification and self-supervised distortion classification [VPMH21]; surgical instrument segmentation and self-supervised stereo disparity estimation [PMVS22]. A review of this topic can be found in [ZWC+22].

2.2.2.5 Meta-Learning

Perhaps the most distinctive of the methods listed is meta-learning, a technique for training models for rapid parameter adaptation from a few labelled examples. The focus is "learning-to-learn"; during meta-training, a variety of tasks are presented within either one of the following optimisation frameworks. Metric-based frameworks approach learning an embedding and similarity/distance metric to compare given queries to the few labelled support examples akin to lookup tables. Model-based focus on the design of network architectures to enable fast adaptation using structures such as memory modules. Lastly, optimisation-based, where the current literature suggests this as the most effective framework in practice, involves learning an initialisation such that within a few gradient steps rapid adaptation of the parameters is achieved. An important example is the Model-Agnostic Meta-Learning (MAML) [FAL17], which is a two-step optimisation process, with an inner loop that updates task-specific parameters and an outer loop that optimises initial parameters for rapid convergence. A notable variant is the Implicit MAML, which decouples the inner and outer loops via implicit differentiation, improving the computational cost [RFKL19]. These optimisation-based meta-learning approaches have shown promise in healthcare translation; for example, fine-tuning a model within a few-shot learning framework for lesion segmentation [KJH+21]. However, as highlighted in recent surveys [PC23]

[RMJ+23], meta-learning in healthcare has received less attention than alternative methods and many significant open challenges remain. I therefore describe metric- and model-based frameworks only briefly, as above.

2.3 Technical Background - Ultrasound

2.3.1 Ultrasound Hardware and Imaging

When thinking of US, most people will imagine black and white feotal imaging. However, a plethora of advanced technologies have been and continue to be developed on top of standard US technology and machines.

A standard feature of US is Doppler flow imaging, which measures the velocity and direction of blood within vessels. This technique operates by measuring frequency shifts in returning echoes, caused by the movement of blood relative to the transducer. There are many variants of Doppler ultrasound techniques. Superb microvascular imaging (SMI) improves the visualisation of low-velocity microvascular flow using a clutter suppression algorithm [ANAS17]. The qualitative evaluation in [ION+17] found that SMI aids in identifying normal tissue, differentiating tumour types (e.g., high- and low-grade gliomas [CXW+23]) and delineating tumour borders [CXW+24]. Other Doppler-based US techniques include functional ultrasound (fUS), which measures changes in cerebral blood volume [DDPT18]. This is achieved using ultrafast ultrasound imaging systems [TF14], which are capable of capturing thousands of frames per second. Similarly to fMRI, fUS can help maximise safe resection by identifying eloquent brain structures [SVS+20].

Another notable advancement in US technology with promising intraoperative and non-invasive applications is focussed ultrasound (FUS). For example, FUS can be used to temporarily open the blood-brain barrier for targeted drug delivery [BH14]. However, the most relevant application for this thesis is the use of high-intensity focussed ultrasound (HIFU) in tumour ablation. HIFU works by concentrating high-intensity ultrasound waves on a specific target, inducing localised thermal ablation (>56°C), which when maintained for a couple of seconds, causes coagulative necrosis and tumour and cancer cell destruction. Studies have shown that this can be achieved both with and without craniotomy [QWK+18]. By eliminating the need for surgical incisions, HIFU minimises damage to surrounding healthy tissue and offers a promising tool for noninvasive neurosurgical interventions.

2.3.2 Ultrasound Registration

A rather important point of exploration for neurosurgical applications is the real-time registration of IOUS images with preoperative images. This can be achieved physically through optical or magnetic trackers [SBP+17]. However, these are often limited to only overlaying information (rigid alignment), without accommodation for deformation and shifting, and the requirement for additional hardware, which can introduce user-based challenges. This motivates the exploration of computer vision-based systems. The first major effort in this direction was the correction of brain shift with the intra-operative ultrasound (CURIOUS) 2018 challenge [XMW+19], where the retrospective evaluation of cerebral tumors (RESECT) [XFU+17] dataset was used for algorithm development. Specifically, the challenge focussed on rigid registration prior to dura opening. This included the winning methodology that used the linear correlation of linear combination similarity metric [WLF+13], and the second-place method, which was based on the self-similarity context (SSC) descriptors [Hei18]. Where both methods achieve a mean Euclidean distance test score of sub 2mm between two corresponding landmark sets, and the recorded mean computational time per case was 20 and 25 seconds, respectively. A recent extension of CURIOUS/RESECT is the brain resection multimodal imaging database (ReMIND) collection [JDK+23]. Although the RESECT dataset included some images after dura opening and after resection, the dataset primarily contains images before dura opening. The ReMIND data critically add post-resection iMRI and IOUS, as well as pre-operation iMRI with pre- and post-dura opening IOUS. This addition of post-resection information is critical for studying the deformation issues created during cutting.

2.3.3 Ultrasound Segmentation

Diving into the deep learning inferno, we encounter the most popular area of interest, which is that of tumour segmentation. Owing to the incredible variability in image quality depending on machine and settings, non-standardisation of the IOUS for brain tumour surgery, difficulty in acquiring ethics, and other factors. The volume of publicly available data is extremely limited which consequently also limits what can be technically achieved. Extending from the

registration subsection, the same group responsible for RESECT later extended the dataset to include tumour segmentations, RESECT-SEG [BCC+22], which is probably currently the most important public dataset. Due to the leveraging of the registration information to support annotation and the consequential opening for future work for the development of methods for dual segmentation and registration, this presents a logical and effective next step. However, the issue with this dataset is the use of 3D US which suffers from inferior imaging quality compared to same-generation, comparable 2D US machines. As highlighted in [CESS+25], the 2D IOUS images captured as part of the project that this thesis was part of are far superior to the RESECT-SEG data.

Various strategies have been proposed to minimise the impact of data scarcity; for example, through the fusion of different data modalities and deep learning training setups. Two of these approaches will be outlined here. In [CESS+25], a segmentation method is proposed using exclusively 2D IOUS images. The approach was based on nnU-NET [IJK+20], a self-configuring U-NET training framework, specifically designed to simplify optimisation for the creation of benchmark / well-performing biomedical image segmentation models. Their model demonstrated strong generalisability across unseen data from different machines, including both images from 2D and 3D acquisitions. A more advanced method is proposed in [DTH+24], where a patient-specific segmentation method is developed that utilises the annotation of pre-operative MRI images to train an MRI-IOUS synthesis Multi-Modal Hierarchical Variational Auto-Encoder model. These synthetic images can then be used to overfit a real-time 2D UNet to achieve patient-specific tumour segmentation.

2.3.4 Thesis Bridge

This thesis is principally concerned with general computer-assisted methods, and relevant topics have been described in this review of the literature. However, since the PhD was also part of the JADS robotic project, methods for assisting the robotic system are also of concern to this work. The issue with this is that although the field of robotic ultrasound has attracted a great deal of attention and output, nothing has been proposed for brain surgery. For example, recent

comprehensive surveys of robotic ultrasound [JSN23] [BJD+24] show not only that no work has been explored for brain surgery. But also, they completely fail to mention it as an avenue for future work. A general background discussion of robotics is beyond the scope of this thesis. Although some robotics work is included in the contributions, no extensive background reading is required. For those interested in robotic US, I recommend those survey papers as a starting point.

As has been discussed, considering the extraordinary acceleration of AI methods and applications, data-driven methods have seen very minimal translation to neurosurgical applications. The direct transfer of principles from other IOUS applications to brain surgery is unlikely due to its unique anatomical and procedural constraints. Key factors include the limited surgical window, the delicate nature of brain tissue, the proximity to eloquent functional areas, and the infiltrative nature of brain tumours. As such, clinical realities are often simplified to facilitate technical feasibility, but this, in turn, distances solutions from real-world surgical needs. These challenges have discouraged research, leading to limited technical developments and a scarcity of available data. As a result, this thesis aims to address fundamental issues that have been explored inadequately.

Part I

Pathological Tissue Delineation

Preamble

The most challenging aspect of IOUS is the interpretation of the US feedback. This is especially the case for brain surgery, where tissue delineation is inherently poor, particularly for pathological tissue. Within the NHS, IOUS training is largely limited to radiologists and sonographers. With surgeons commonly learning "on the job" and relying on general US training or very limited IOUS training. This is definitely the case within the neurosurgical community, where adoption of IOUS as a regular tool has been notably slow. Much of the reason for this is the perceived difficulty in mastering IOUS. Exacerbated by the wide variety in US machine quality and the limited literature on the topic. Up until the publication of the first chapter, there were no public studies measuring the issue of interobserver variability among experienced neurosurgeons in IOUS. The main objective of the study is to highlight IOUS operator challenges and the need for more formal training. In addition, through studying interobserver variability, the study aims to indirectly measure the more fundamental challenge of tumour margin aleatoric uncertainty; given the resolution of current state-of-the-art US machines. In light of the conclusion of the study, an annotation method is proposed to directly address the issue of uncertainty over tumour margins through Gaussian distributions representing confidence. This annotation method has been shown to offer superior downstream application for data learning, particularly when data is limited, as the annotations themselves are better representative of the underlying information of the pathological tissue.

Within the context of this thesis, this outcome of this research is fundamental for developing topic appropriate automatic tissue characterisation algorithms, which can be leveraged within a human-in-the-loop framework (assisted scanning) or to drive a fully autonomous robotic system.

Chapter 3

Interobserver Variance Of Tumour Annotation

3.1 Introduction

Maximal-safe resection of brain tumours is a key pillar of modern neuro-oncology management, improving symptoms, quality of life and overall survival [HJB16]. Accurate delineation of a lesion from surrounding normal functional brain tissue remains challenging but is critical to ensure optimal resection. Ultrasound (US) has been employed intraoperatively (intraoperative ultrasound IOUS) in neurosurgery for 70 years [FWN50]. Applied for surgical guidance, as it allows real-time detection, characterisation and outlining of tumours. Unlike Magnetic resonance imaging (MRI) - specifically intraoperative MRI (iMRI) - it integrates easily into the surgical workflow and is relatively affordable. With the availability of high-quality US machines and navigated US co-registered with preoperative MRI and computed tomography (CT) [DR01, HDIT97, JRHP12, OKN+10, MMQ+16], its adoption has increased over the last two decades.

However, factors associated with the quality of the modality as well as standardisation of training has effected universal adoption [DLGS+22]. There remains a perceived steep learning curve secondary to limited fields of view with unfamiliar topographical representation, artefacts, the

unique visuo-tactile task and the difficulty with gaining experience outside of the intraoperative setting. There are also concerns regarding the imaging accuracy and granularity. These factors can make learning IOUS difficult, with the potential to impair tumour and tumour boundary detection with the risk of leaving unintended residuum or causing inadvertent damage. This challenge is further compounded by the inherent great variation in types of brain lesions, their appearances, the degree of infiltration and the intraoperative changes (such as oedema and haemorrhage) which can further confound.

Presently, there have been only a few meta-analyses looking at the impact of IOUS in glioma surgery. One pooled series reported an average 77% gross total resection rate in 739 patients undergoing IOUS-guided resection (71.9% in high grade glioma, HGG, compared to 78.1% in low grade glioma, LGG) which was comparable to other forms of navigation. A recent meta-analysis of 409 diffuse gliomas compared the accuracy of IOUS to the reference standard post-operative MRI. They found that IOUS was an effective technique in assessing diffuse glioma resection, with a 72.2% pooled sensitivity and a 93.5% pooled specificity [TBT+19]. Whilst these results are encouraging the current evidence supports a need to improve IOUS accuracy if it is to become part of the standard of care in brain tumour surgery. Trials assessing the role of US in neurosurgery, such as the randomized controlled trial Functional and Ultrasound-Guided Resection of Glioblastoma (FUTURE-GB), and refining of US techniques are therefore needed [PCC+22].

There are three aims of this study. Firstly, to assess whether tumour boundary detection on IOUS is challenging, we measure interobserver variation between regular IOUS operators in segmenting US images of brain lesions. Secondly, we model the pixel intensities of the segmented tumour boundaries to mathematically model the clarity and blurriness - factors relating to issues of low resolution and low signal-to-noise ratio - of tumour boundaries. Finally, we evaluate the interobserver variation of bounding boxes to assess whether this annotation method has a role as an alternative, complimentary simplified method for outlining lesion margins. Additional evaluation is performed to determine if these bounding boxes can be used as a guide to improve the accuracy of the segmentation.

To improve the utility of IOUS in neurosurgery, the understanding of the limitations of tumour-margins delineation capabilities needs to be understood and measured. With standardised training, new supporting techniques as well as the development of new tools required to reduce uncertainty and error. By formally addressing the issue of tumour segmentation error and uncertainty, we hope to highlight a fundamental challenge with IOUS that has hindered universal adoption. Until now, this issue has largely been an implicit challenge recognised by experienced IOUS operators.

3.2 Materials and Methods

A preliminary study was conducted using 4 annotators (AN) experienced with IOUS - a neuroradiologist and 3 neurosurgeons - to determine the foundation of our hypothesis. The neuroradiologist and the three neurosurgeons are all post-training doctors. All clinicians involved have extensive research backgrounds and familiarity with segmentation tools and protocols. Specifically: Neuroradiologist - 10 years of US and neuroimaging experience; An1 - 10 years of neuro-oncology experience with 10 years of $\approx$ 2-3 cases per week using IOUS; An2 - 8 years of neuro-oncology experience with 6 years of $\approx$ 1 case per week using IOUS; An3 - 9 years of neuro-oncology experience with 9 years of $\approx$ 2 cases per week using IOUS. The order of the annotators has been randomised.

3.2.1 Data

The dataset consists of 30 images, from 30 patients, taken during brain surgery at Imperial College NHS Trust London. The images were retrospectively selected by the neuroradiologist from cine clips of US sweeps that captured the entire tumour. Images with a field of view which covered the boundaries of the tumour and at least 2cm of surrounding normal brain were selected with reference to the preoperative MRI (Magnetom Vera 3T, Siemens) to ensure accuracy. All images were captured before tumour resection. For patient information please see Tab. 3.1. Images were acquired using a Canon i900 US machine (Canon Medical Systems, Japan) with the 8MHz i8MCX1 microconvex probe, and were stored using the Digital Imaging and Communications in Medicine (DICOM) format. For all images, the dynamic range was fixed at 70dB. The aperture was set to either 1 or 3. The power level was set to either 5 or 7. The gain varied between 79 to 98. The maximum depth was varied between 5cm to 11cm, with the depth focus between 1.8cm to 8.1cm. All images are of size 960×1280.

The study had full local ethical approval by the HRA and Health and Care Research Wales (HCRW) authorities. Study title - US-CNS: Multiparametric Advanced Ultrasound Imaging of the Central Nervous System Intraoperatively and Through Gaps in the Bone, IRAS project

Table 3.1: Patient information. KPS = Karnofsky Performance Status, ASA = American Society of Anesthesiologists Grade, SCC = Squamous Cell Carcinoma, SS = Supratentorial Subependymoma, DNET = Dysembryoplastic Neuroepithelial Tumor, TB = Tuberculosis, Lt = Left, Rt = Right.

ID	G	Age	Lesion	Location	Histology	KPS	ASA
1	F	60-69	Metastasis	Rt Parietal	SCC - lung	80	3
2	M	30-39	Glioma	Rt Frontal	Astrocytoma	90	1
3	M	50-59	Glioma	Rt Parietal	Glioblastoma	80	2
4	F	40-49	Glioma	Lt Frontal, Parietal	Astrocytoma	90	2
5	M	20-29	Glioma	Lt Parietal	Astrocytoma	80	2
6	M	60-69	Glioma	Lt Parietal	Glioblastoma	70	2
7	M	50-59	Glioma	Lt Frontal	Oligodendroglioma	90	3
8	M	50-59	Glioma	Rt Frontal	Glioblastoma	90	2
9	M	60-69	Glioma	Lt Parietal	Glioblastoma	90	3
10	M	50-59	Glioma	Rt Frontal	Glioblastoma	90	2
11	M	40-49	Glioma	Lt Parietal	Glioblastoma	80	2
12	M	60-69	Glioma	Rt Parietal	Glioblastoma	70	3E
13	M	70-79	Glioma	Rt Frontal	Glioblastoma	70	2
14	M	10-19	Granuloma	Lt Temporal	Granuloma TB	100	1
15	F	60-69	Glioma	Rt Parietal	Glioblastoma	70	2
16	M	60-69	Glioma	Rt Parietal	Glioblastoma	80	2
17	M	40-49	Glioma	Lt Temporal	Glioblastoma	100	2
18	F	60-69	Glioma	Lt Temporal	Glioblastoma	70	2
19	M	40-49	Glioma	Lt Frontal	Glioblastoma	80	2
20	F	50-59	Ependymal	Lt Temporal	SS	90	2
21	M	30-39	Glioma	Post Fossa	Pilocytic astrocytoma	90	1
22	M	60-69	Glioma	Rt Parietal, Occipital	Glioblastoma	70	2
23	F	60-69	Glioma	Lt Frontal	Oligodendroglioma	90	2
24	M	30-39	Glioma	Lt Parietal	Glioblastoma	90	1
25	M	70-79	Glioma	Rt Parietal	Oligodendroglioma	90	2
26	F	20-29	Glioma	Rt Frontal	Astrocytoma	90	2
27	M	40-49	Glioma	Rt Frontal	Astrocytoma	90	2
28	F	10-19	Glioneuronal	Lt Frontal	DNET	90	1
29	F	40-49	Glioma	Lt Temporal, Occipital	Oligodendroglioma	90	2
30	F	30-39	Glioma	Lt Parietal	Astrocytoma	90	2

ID: 275556, Protocol number: 22CX7609, REC reference: 22/WA/0259, Sponsor: Research Governance and Integrity Team (RGIT).

3.2.2 Annotation protocol

The boundaries of tumours were annotated, for both segmentations and bounding boxes, using 3D Slicer (5.4.0) [FBKC+12]. The annotations provided by the neuroradiologist were made using the benefit of the full US dataset, cross-registration with the preoperative MRI and patient metadata; and we define these as the reference standard/ground truth for this study. The other annotators were given only the individual 2D US images – these annotations were used to evaluate the consistency and accuracy of the annotation capability, comparing inter-neurosurgeon variation and dissimilarity against the annotations of the neuroradiologist. Whilst this is unlike normal clinical practice (where the MRI and full real-time US would be available and employed by the operator) the aim of this study is to assess the ability of B-mode US to delineate brain lesion margins in isolation. The utility of bounding boxes as a guide to refine tumour boundary segmentation was then assessed by An1 repeating their segmentations 3 months later (to mitigate bias) with the reference standard bounding boxes produced by the neuroradiologist overlaid. The bounding box annotations are created by taking the maximum and minimum, x and y coordinates from the annotators' segments, fitting a box around the outer limits of the segment - as opposed to re-annotating.

3.2.3 Statistical analysis

Three statistical metrics were used to compare the similarity between annotations. Within the mathematical framework, let A and B represent two annotations from the same image i.e. A could be the neuroradiologist's annotation and B, one of the neurosurgeon's annotations. The first two metrics that we define are methods to quantify the degree of overlap - indirectly the degree of similarity in the shape and volume - between A and B. The intersection over union (IoU) [RTG+19] $IoU = \frac{|A \cap B|}{|A \cup B|}$, which is defined as the set of image coordinates occupied in

the intersection of the two segmentations, divided by the union of the two segmentations, the total set of image coordinates occupied by both segmentations. The Sørensen-Dice Similarity Coefficient (DSC) [HPP+20] $DSC = \frac{2|A \cap B|}{|A|+|B|}$, which is defined as twice the intersection of the two segmentations divided by the cardinalities of the two segmentations. For both IoU and DSC, the closer the value to 1 the better the score.

For evaluation of the uncertainty of the boundary delineation, the contours of the segmentations were compared using the Hausdorff Distance (HD) [KS19] $\widehat{HD}(A_c, B_c) = \max_{a_c \in A_c} \min_{b_c \in B_c} ||a_c - b_c||$, $HD = max(\widehat{HD}(A_c, B_c), \widehat{HD}(B_c, A_c))$, which determines the maximum Euclidean distance ($pixels$) between all closest point pairs between the two contours sets - the subscript c denotes the contours of the corresponding segmentations, where the contours are the lines intersecting the endpoints of the segmentations. For this metric, the closer the value is to 0 the better.

This analysis was performed using Python, pynrrd [EEBJ+19] for reading the segmentation files and scikit-image [VDWSN+14] and SciPy [VGO+20] for the quantitative analysis.

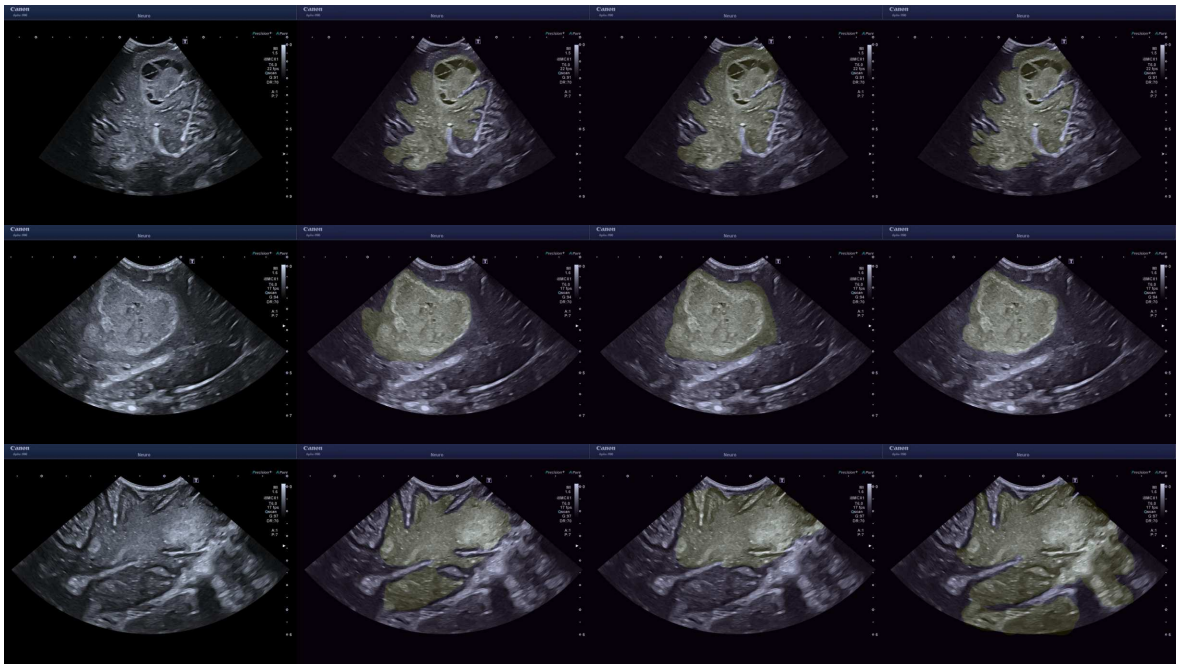
3.2.4 Tumour Border Pixel Dispersion Substudy

To assess the visible distinguishability of the tumour from the surrounding normal tissue, the dispersion of pixel intensities around the neuroradiologist's tumour margins are evaluated. By measuring the pixels along the segmentation's contour plus a 10 pixels border, perpendicular to, and on both sides of the contour. To calculate the properties of the distributions, a local maxima peak-finding method is implemented. Where one peak = unimodal, two = bimodal and three or more = other (could be either multimodal or uniform).

3.3 Results

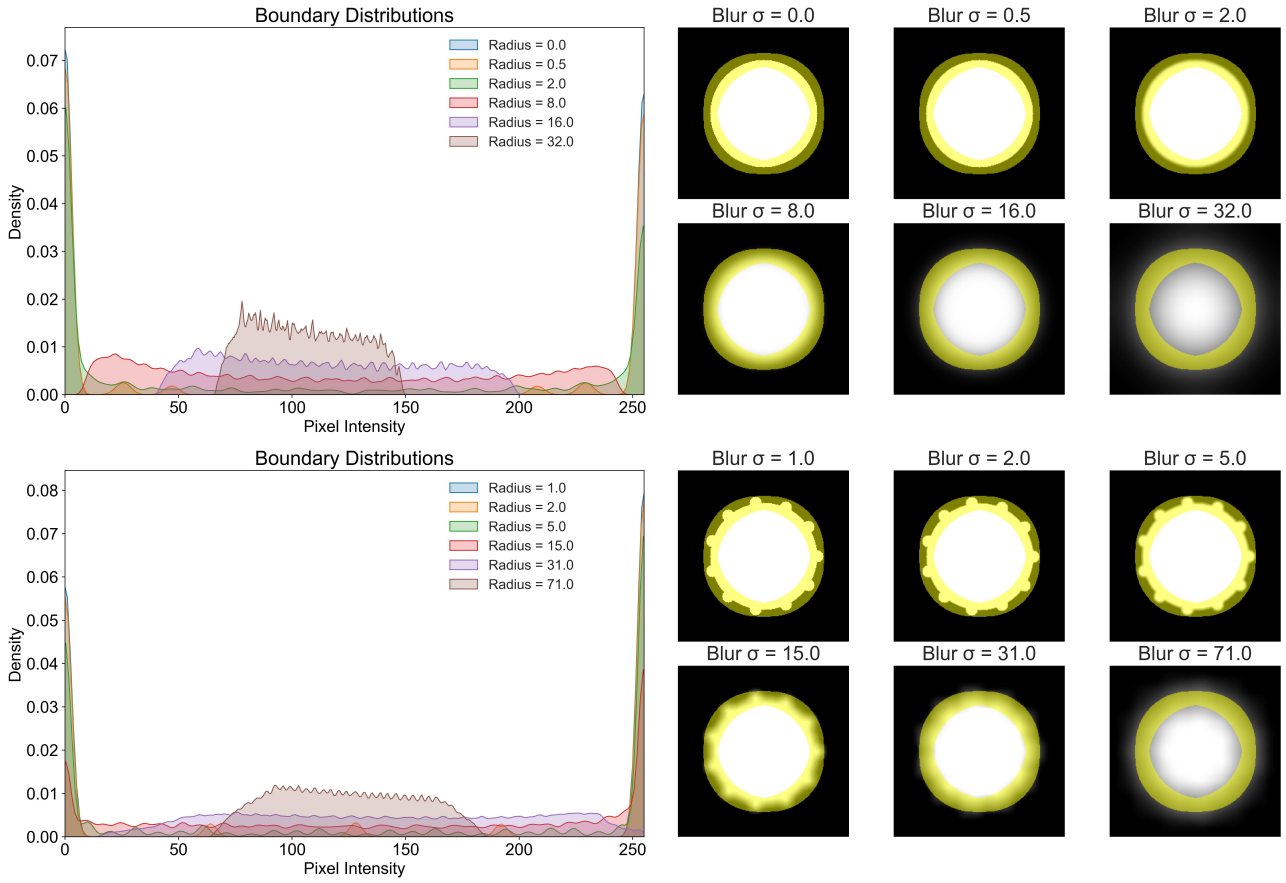
3.3.1 Inter-Neurosurgeon annotation variance

Figure 3.1: Inter neurosurgeon annotation variance. From top to bottom - ID 14, ID 16, ID 18. ID is the image index.



All results are presented in the following format - Metric[Mean:interquartile range (IQR)]. The similarity measure between the neurosurgeons' annotations is as follows - IoU[0.789:0.115], DSC[0.876:0.072], HD[103.227:73.071]. Although the overlap scores indicate general similarity between the annotations, there are still inconsistencies, which when considering the precision required for tumour resection, this marginal difference can be considered impactful. The average HD, on the other hand, is a significant result. Showing that there is frequent disagreement on at least one point along the tumour boundary. Figure 3.1 highlights 3 examples of the annotation variance. All figures show cases with minor and severe variance, to visually highlight the features and general causes of uncertainty.

Figure 3.2: The top row/example shows the effect on the boundary of a binary circle when Gaussian blurring is applied - using a kernel density estimate plot. The bottom row/example shows the same but using a box filter, where the circle also now has protrusions, unaffecting the defined boundary represented by the yellow mask.

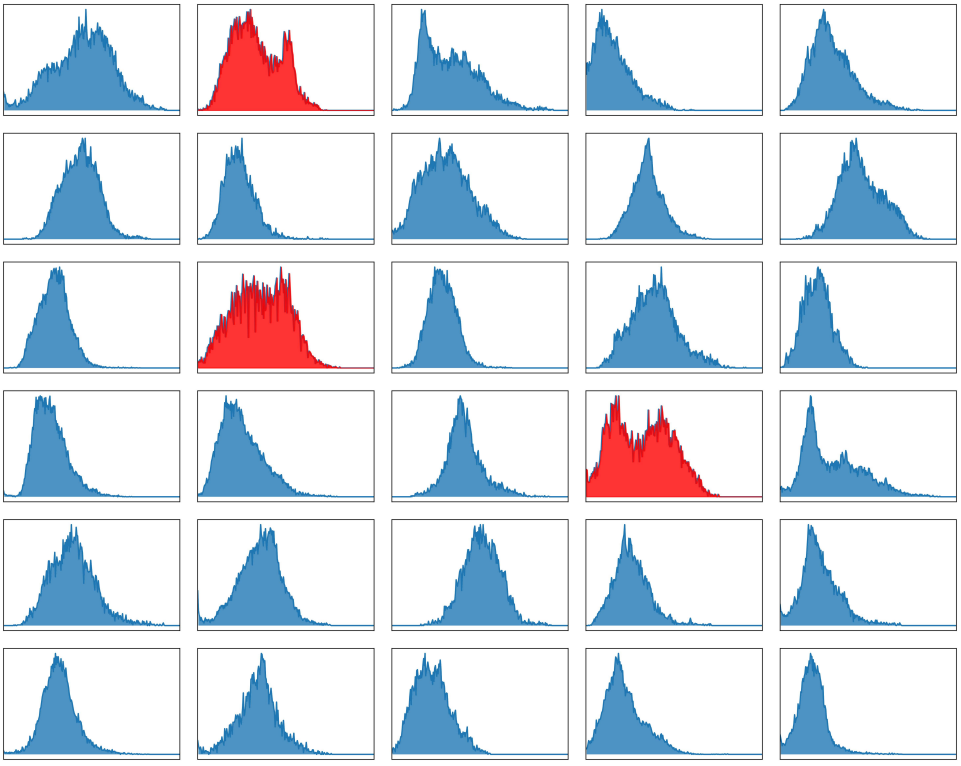


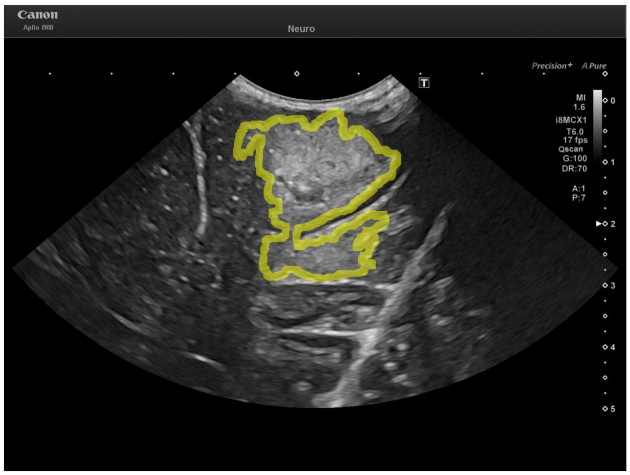
3.3.2 Tumour Border Pixel Dispersion Substudy

First, we provide a simplified example to illustrate how smoothing affects pixel distributions along a binary boundary (perfect separability). In Figure 3.2 a binary circle is created and both Gaussian and box filters are applied with increasing amplification. For the case of the binary circle without smoothing, the boundary is perfectly defined and as such the pixel distribution is bimodal. However, as the smoothing factor increases, the distribution tends towards uniformity or multimodal and under severe blurring, becomes unimodal.

From evaluating the tumour boundaries, from the 30 images, 27 were classified as unimodal and 3 as other Figure 3.3. This result is strong evidence of the severity of the pixel intensity variation along the tumour boundary, which provides mathematical evidence for the ambiguity in defining the discrete, definite boundary points.

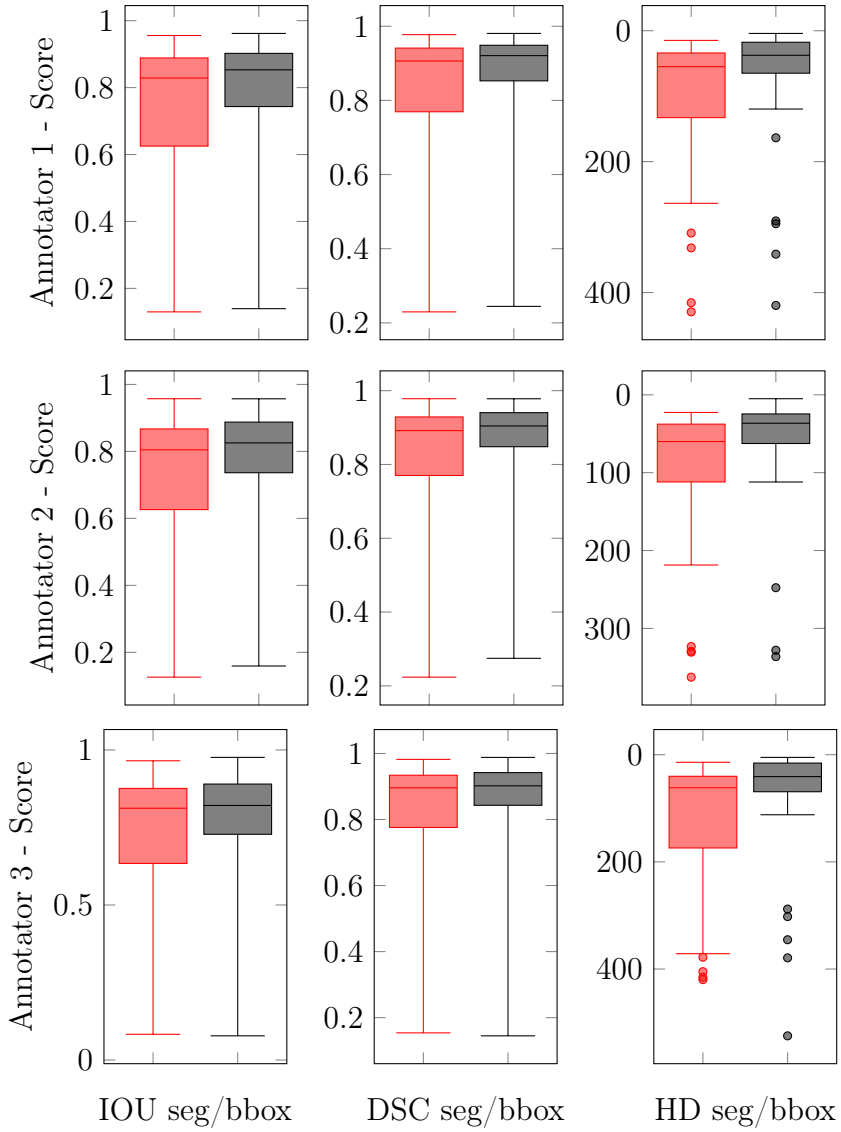
Figure 3.3: Top is a plot of all tumour boundary pixel intensities distributions. Blue highlights unimodal distributions and red highlights other. Bottom is an example of tumour boundary is shown using ID 009. ID is the image index.





3.3.3 Neuroradiologist-Neurosurgeon annotation variance

Figure 3.4: Box and Whisker plots of the annotators IoU, DSC and HD scores on the 30 images - red = segmentation, black = bounding box. What is highlighted is the greatly reduced median score and IQR when using the bounding box method. IoU is Intersection Over Union, DSC is Sørensen-Dice Similarity Coefficient, HD is Hausdorff Distance and IQR is Interquartile Range.



The results of the similarity comparison between the reference standard annotations produced by the neuroradiologist (which benefited from correlation with the preoperative MRI and the full US dataset) and the neurosurgeons is tabulated in Tab. 3.2,3.3, and visualised using a box and whiskers plot in Figure 3.4. This showed a moderate interobserver variance between the reference standard segmentations' and the annotations performed by the neurosurgeon on the single slice B-mode images alone, highlighting the potential limitations and uncertainties of

Table 3.2: Shown are the IoU and DSC similarity results. The closer the value to 1 the more similar the annotations are. IoU is Intersection Over Union, DSC is Sørensen-Dice Similarity Coefficient.

	An1				An2				An3			
	Seg		BBox		Seg		BBox		Seg		BBox	
Image	IoU	DSC	IoU	DSC	IoU	DSC	IoU	DSC	IoU	DSC	IoU	DSC
1	0.952	0.976	0.954	0.976	0.875	0.933	0.862	0.926	0.941	0.970	0.971	0.985
2	0.908	0.952	0.962	0.981	0.911	0.954	0.957	0.978	0.880	0.936	0.847	0.917
3	0.888	0.941	0.907	0.951	0.906	0.951	0.914	0.955	0.891	0.943	0.955	0.977
4	0.902	0.948	0.926	0.962	0.848	0.918	0.886	0.940	0.389	0.560	0.230	0.374
5	0.905	0.950	0.940	0.969	0.856	0.923	0.912	0.954	0.893	0.943	0.856	0.922
6	0.570	0.726	0.536	0.698	0.811	0.895	0.854	0.921	0.876	0.934	0.893	0.943
7	0.578	0.733	0.734	0.846	0.781	0.877	0.916	0.956	0.819	0.901	0.940	0.969
8	0.889	0.941	0.863	0.927	0.957	0.978	0.954	0.977	0.965	0.982	0.971	0.985
9	0.591	0.743	0.830	0.907	0.693	0.819	0.746	0.854	0.679	0.809	0.724	0.840
10	0.902	0.948	0.960	0.980	0.788	0.881	0.733	0.846	0.913	0.955	0.943	0.971
11	0.644	0.783	0.581	0.735	0.800	0.889	0.919	0.958	0.670	0.803	0.777	0.875
12	0.833	0.909	0.776	0.874	0.872	0.932	0.887	0.940	0.923	0.960	0.976	0.988
13	0.823	0.903	0.816	0.899	0.809	0.895	0.843	0.915	0.865	0.928	0.870	0.931
14	0.265	0.419	0.217	0.357	0.219	0.359	0.199	0.332	0.206	0.342	0.199	0.333
15	0.130	0.230	0.148	0.257	0.125	0.221	0.159	0.275	0.129	0.228	0.143	0.251
16	0.885	0.939	0.903	0.949	0.809	0.894	0.855	0.922	0.802	0.890	0.763	0.865
17	0.706	0.828	0.702	0.825	0.548	0.708	0.649	0.787	0.616	0.762	0.780	0.876
18	0.168	0.287	0.139	0.245	0.167	0.286	0.172	0.293	0.083	0.154	0.078	0.145
19	0.901	0.948	0.875	0.933	0.790	0.883	0.805	0.892	0.835	0.910	0.616	0.762
20	0.956	0.977	0.943	0.971	0.870	0.930	0.819	0.901	0.839	0.912	0.747	0.855
21	0.873	0.932	0.902	0.948	0.452	0.623	0.475	0.644	0.853	0.920	0.881	0.937
22	0.849	0.918	0.850	0.919	0.854	0.921	0.888	0.940	0.876	0.934	0.877	0.934
23	0.402	0.573	0.208	0.345	0.624	0.769	0.746	0.854	0.534	0.696	0.233	0.378
24	0.743	0.853	0.785	0.880	0.632	0.775	0.723	0.839	0.694	0.820	0.740	0.850
25	0.838	0.912	0.883	0.938	0.904	0.950	0.919	0.958	0.751	0.858	0.744	0.853
26	0.619	0.765	0.773	0.872	0.561	0.718	0.698	0.822	0.572	0.728	0.707	0.829
27	0.873	0.932	0.857	0.923	0.813	0.897	0.764	0.866	0.804	0.891	0.836	0.911
28	0.803	0.891	0.874	0.933	0.773	0.872	0.832	0.908	0.756	0.861	0.908	0.952
29	0.791	0.883	0.867	0.929	0.620	0.766	0.754	0.859	0.622	0.767	0.807	0.893
30	0.821	0.902	0.810	0.895	0.878	0.935	0.816	0.899	0.832	0.908	0.879	0.936
Mean	0.734	0.821	0.751	0.827	0.719	0.812	0.755	0.837	0.717	0.807	0.730	0.808

Table 3.3: Shown are the HD similarity results. The closer the value to 0 pixels the better aligned the annotation margins are. HD is Hausdorff Distance.

	An1		An2		An3	
Image	Seg	BBox	Seg	BBox	Seg	BBox
1	14.765	8.544	43.000	36.620	14.142	5.000
2	22.204	4.123	25.000	5.000	52.802	46.271
3	33.734	16.492	32.558	14.142	23.324	9.220
4	29.069	8.246	42.438	33.061	371.389	379.120
5	51.614	13.153	57.245	19.416	40.162	23.345
6	162.926	163.515	56.223	35.511	52.924	17.804
7	74.632	40.311	32.650	7.280	42.579	6.000
8	52.000	48.010	33.121	7.616	28.636	6.403
9	137.773	49.477	74.686	47.265	79.202	48.836
10	25.807	4.243	55.803	44.598	18.682	6.403
11	117.652	119.620	362.627	15.000	138.105	69.231
12	66.000	70.000	68.447	47.170	43.600	7.071
13	46.174	28.844	62.000	30.000	40.522	23.324
14	331.724	341.264	329.524	336.265	331.965	345.307
15	263.610	294.703	218.563	247.746	254.342	288.043
16	36.770	36.000	81.006	44.000	90.554	68.710
17	309.015	110.000	330.510	102.000	377.922	62.169
18	429.439	419.640	323.303	328.056	404.901	524.675
19	27.295	23.000	58.052	36.401	419.715	112.058
20	30.887	23.000	64.938	58.694	87.727	84.000
21	42.012	15.000	119.549	111.973	33.377	14.866
22	68.447	28.425	70.264	25.000	43.600	24.000
23	415.473	290.493	57.079	44.721	415.120	302.154
24	71.007	48.703	89.022	63.789	80.000	53.488
25	56.511	20.809	22.627	14.213	63.600	35.468
26	156.259	45.188	171.047	70.767	175.026	57.271
27	15.000	10.440	24.000	24.413	29.547	21.024
28	34.670	30.067	36.056	30.017	35.228	13.601
29	53.254	38.949	180.049	79.649	171.234	54.406
30	67.231	40.792	26.926	27.731	59.641	24.739
Mean	108.098	79.702	104.94	66.270	133.986	91.134

isolated B-mode in defining tumour boundaries.

With conversion of the segments to bounding boxes there was a noticeable improvement in annotation similarity with a sizable decrease in the IQR. For all annotators, the conversion to a bounding box, overall, increases the annotation similarity. The decrease in the IQR, as a percentage, for [IoU, DSC, HD] are as follows - An1=[40%, 44%, 52%], An2=[37%, 42%, 49%], An3=[33%, 37%, 60%].

Further evaluation of the bounding box proposal is conducted by measuring individually for each image, the percentage of the neuroradiologist's segmentation contained within the neurosurgeon's corresponding bounding box. The average results per annotator are - An1:98.387%, An2:98.833%, An3:99.052%. From the results, it can be concluded that the bounding box approach is usable for localising the entire tumour mass, and is a suitable method for reducing inter-observer annotation variance.

Sample images are displayed in Figure 3.5,3.6. An assortment of different sources of boundary aleatoric uncertainty are shown in Figure 3.5. Including images containing fuzzy borders and continued hyperechogenicity extending beyond the tumour. Shown in Figure 3.6 are example cases where using bounding boxes has substantially improved the annotation similarity.

3.3.4 Improving segmentation accuracy using overlaid bounding boxes as a guide

In the substudy looking at the impact of overlaid reference bounding boxes on An1 segmentation accuracy, we found a substantial improvement in segmentation similarity between the neurosurgeon and reference standard. The previous results compared to the new results are as follows - without bounding box [IoU:0.734, DSC:0.821, HD:108.098], with bounding box [IoU:0.858, DSC:0.922, HD:52.072]. This result shows that the bounding box can be used as a visual anchor to minimise the uncertainty when segmenting.

Figure 3.5: Uncertainty caused by unclear boundary. The left images are the original images, the middle from the neuroradiologist and the right from a neurosurgeon. From top to bottom - ID 13 An3, ID 07 An3, ID 18 An2, ID 19 An2. ID is the image index.

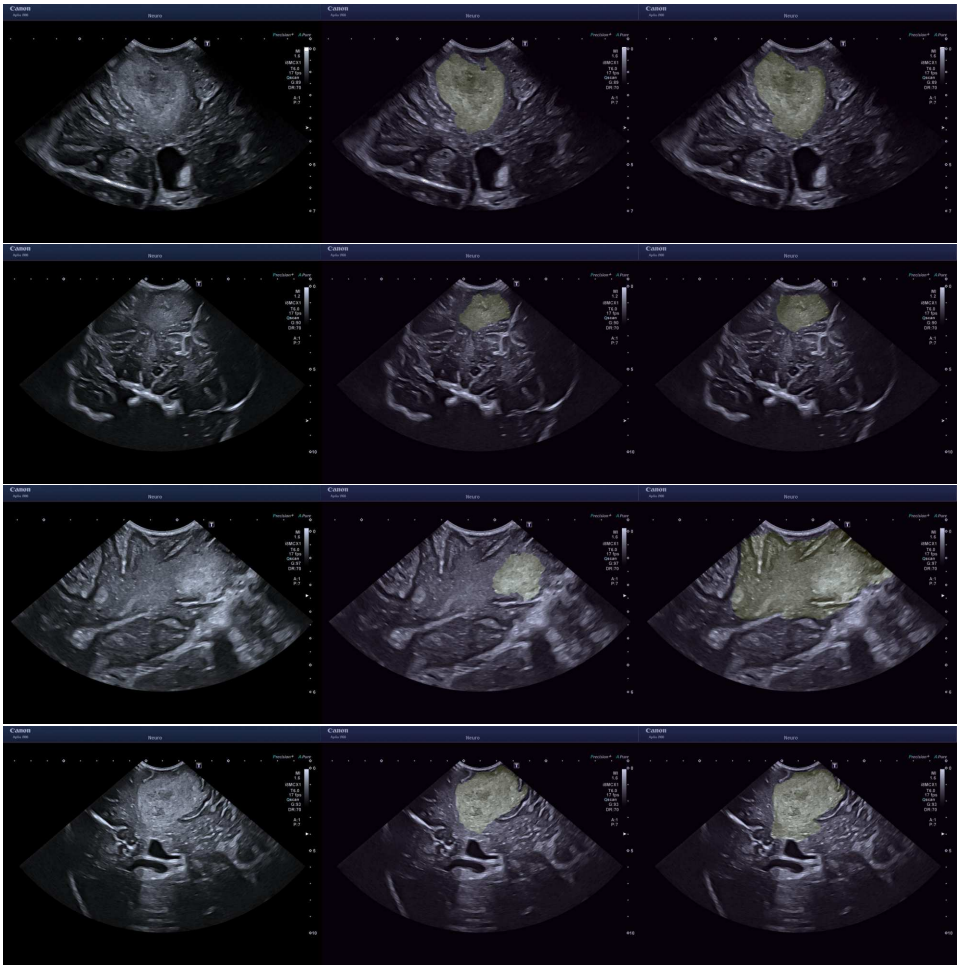
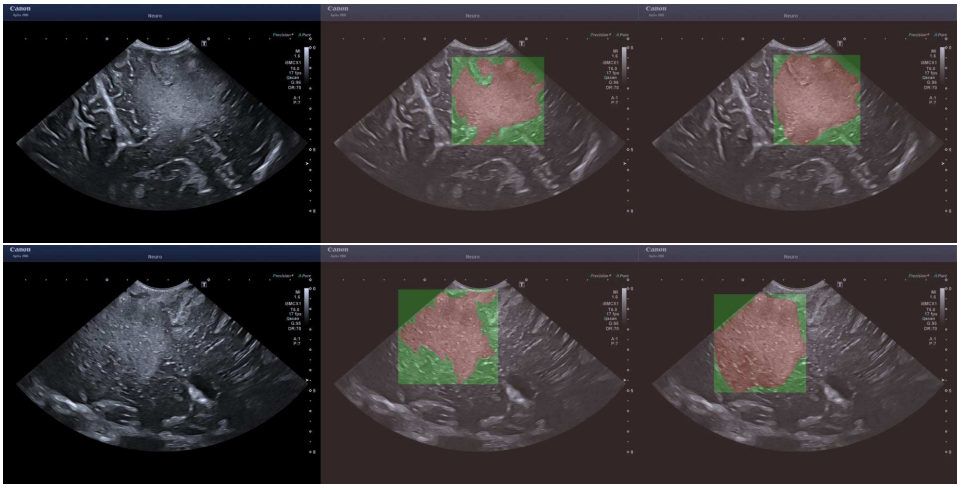


Figure 3.6: Examples for where the segments have large dissimilarity whilst the bounding boxes do not. The left images are the original images, the middle from the neuroradiologist and the right from a neurosurgeon. From top to bottom - ID 11 AN 2, the bottom from ID 26 AN 1. ID is the image index.



3.4 Discussion

3.4.1 Historical background and clinical landscape

Several intraoperative imaging devices have been developed to help neurosurgeons localizing the tumour during surgery. Aside from CT scan and MRI scans, introduced in clinical practice during the '70s and in the '80s, neuro-navigation frameless devices were introduced during the '90s, and they were the first tools allowing intraoperative localization of the tumour [Enc09], although they rely on pre-operative images rather than real time acquisition. The challenging problem of having a real time tool for tumour location and visualization has been partially sorted by the gradual introduction of IOUS in the armoury of neurosurgical equipment [DR01]. Interestingly, in several cases the signal obtained from IOUS was different when compared to that obtained using traditional imaging technologies (CT or MRI scan), thus sparking attention on the fact that the IOUS could be used not only as an intraoperative aid, but also as a complementary tool for those lesions of unclear nature or margins [DR01]. Moreover, coupling of the IOUS with MRI based neuro-navigation has greatly improved the possibility of midline shift adjustment during surgery [LTR+06, LKO+03, OKN+10] and has been proven useful to improve tumour demarcation when close to eloquent areas [RLR+07]. In more recent years, further technological advancement has led to the development of 3D IOUS [BJT+21, SZP+22, UOM+02, UL19]. The 3D reconstruction is automatically generated by the neuro-navigation software after an intraoperative acquisition through a single spatial plane. The images can be integrated with doppler angiography when required, so that vessel encasement by an intracranial mass or aneurysm can also be detected [ŠBT+22, PMM+15]. Recent research has also been focusing on the possibility of integrating IOUS with contrast [ACM+16, ZSX+23].

3.4.2 Findings, challenges, and future perspectives

In this study, we highlight the specific challenge of tumour detection and tumour boundary delineation in cranial IOUS. Here we demonstrate that there remains moderate to high interob-

server variation in the identification and segmentation of tumours on B-mode images acquired on a modern, current-generation US scanner, between four individuals with experience in IOUS-guided brain surgery. Three broad elements likely contribute to this variance, 1) the specific qualities of brain tumours, 2) technical ultrasound factors and 3) operator influences.

Firstly, there are inherent features of brain tumours themselves which can make them difficult to delineate. Gliomas, in particular, are well-known to be infiltrative tumours, meaning that they spread cancer cells beyond their obvious radiological margins [PAB+14, SCJ+12]. Moreover, both high-grade gliomas and low-grade gliomas often show a degree of surrounding oedema which is, to this day, challenging to interpret in terms of differential diagnosis between reactive inflammatory tissue or actually infiltrated brain, even using more established neuroimaging tools such as MRI. Moreover, the intraoperative imaging quality varies based on the stage of surgical resection [MMQ+16, UL19]. Superficial, small sized lesions are typically very well visualized by the acquisition and during resection. However, even modest to moderate amount of bleeding causes a visible artifact that can hamper surgical view beyond the limits of resection. To detect residual tumour, it is crucial to perform accurate haemostasis and remove all haemostatic material to the surgical cavity.

One of the main points of discrepancy in the present series was the definition of tumour margins. For example, in some cases, the lack of clarity over whether the highlighted hyperechogenicity of the tissue was caused by a continuation of the tumour mass or just reactive oedema, introduces significant heteroscedastic aleatoric uncertainty when defining the exact margin. This challenge is further emphasised by how there remains no reference standard imaging technique that absolutely defines tumour extent, nor is it usually possible to remove tumours en bloc in intra-axial neurosurgery, precluding accurate histological correlation of tumour margins.

Secondly, there are unique challenges that US presents. There are numerous ways that an ultrasound image can be altered, including changes in settings (such as gain and frequency), ultrasound machine, probe type, probe contact and probe angle. In most cases, several of these parameters need to be intentionally tailored to the particular tumour being imaged. For instance, using a low-frequency probe, which has a trade-off in reduced spatial resolution, to

visualise a deep tumour versus a high-frequency probe, with high spatial resolution, to image a superficial cortical tumour. This wide range of US options can greatly alter the final image creating another source of variance and in turn, aleatoric uncertainty which is arguably greater than typically seen with other established imaging modalities such as CT and MRI.

US is also vulnerable to several unique artefacts, such as acoustic shadowing [WDA+23b] and acoustic enhancement, which can alter and obscure the image creating a further source of uncertainty. These issues could be mitigated somewhat by the establishment of a standardised protocol for US settings and image acquisition. However, even then, it would be impossible to fully account for all scenarios due to the wide spectrum of tumours and anatomical locations. The potential for confounding artefacts and uncertainty regarding tumour boundaries further increases as surgery progresses due to increased deformation, oedema and potential obscuring blood products. This uncertainty could also be reduced with operator experience, which links to operator factors which are the final source of variance. Currently, neurosurgical training in US is predominantly experiential based on exposure to live cases in theatres. Whilst this is an essential aspect of learning a new surgical skill this can greatly prolong the learning phase due to a relatively low rate of exposure to the imaging technique as it necessitates an intraoperative setting with a craniotomy window. This is in contrast to CT and MRI, which neurosurgeons are much more comfortable with interpreting, owing, in part, to these being readily available and performed regularly on most patients.

There are several complementary ways that IOUS accuracy could be improved and the steep learning curve could be flattened and shortened. This includes the use of advanced multimodal US (including contrast-enhanced US) and navigated US. In addition, dedicated courses employing US phantoms with brain tumour models can greatly help, by providing both applied formal training in US theory in addition to hands-on time with US scanning. In all cases, however, these approaches still require significant time investment and training. In this context, automated, computer-assisted, detection and segmentation of brain tumours on ultrasound would be a desirable and useful additional tool. Automated segmentation of brain tumours on MRI has been well explored with several robust and open source tools now available [PBEe22]. In contrast, development in automated segmentation of IOUS images is in its infancy with no

applications yet available. There are several reasons for this. As illustrated in this study it is challenging to establish a ground truth dataset using manual segmentation due to high inter-observer variance - especially as most published data usually contains only individual frames or volumes. Furthermore, unlike MRI there is a paucity of neuro-oncology IOUS imaging datasets. This is likely due to the relatively low number of IOUS scans acquired, the greater logistical challenges with saving and downloading scans, the potentially large file sizes of video acquisitions, the greater variation of IOUS across sites and the need for highly experienced annotators to perform the time-consuming segmentations. Considering these many boundaries, we are far off the realisation of a reliable system to rapidly automatically and accurately segment IOUS brain tumour images. To bridge this gap, here we assess the utility of bounding boxes as an additional complementary tool for simplified tumour detection and delineation.

Unsurprisingly, we found much lower interobserver variation when using bounding boxes to define tumour location and margins compared to segmentations. The advantage of our proposed use of bounding boxes is a step in the direction of overcoming the above-mentioned challenges. From a clinical perspective, the bounding box would be useful for training purposes and immediate identification of the tumour mass. While this system is expected to be lower in specificity, the high sensitivity should assist inexperienced surgeons in detecting tumours and providing an area to focus on. Further, the whole signal change (fuzzy margin) would be included in the bounding box, thus making sure that there is no missing tumour from the targeted area. The process of annotation is also much quicker for bounding boxes although this will be annotator-dependent, from our experience, bounding box annotation may take as little as 1/3 of the time of segmentation, reducing the manual labour cost.

In computer vision and AI, segmentation methods will typically define, assume that the segmentation task can be framed as separating an image into sub-regions with defined and complete boundaries. For example, explicitly through the definition of a mathematical optimisation framework, or implicitly by training a neural network on common datasets. Because of this, the task of semantically segmenting brain tumours in US becomes uniquely difficult. From an engineering perspective, there are a large number of benefits to using bounding boxes. First and foremost, the reduced annotation complexity should facilitate the collection of large datasets.

Accelerating technical and clinical research into this topic. For technical development, the primary benefit is the reduced complexity of the estimation task, which should lead to highly accurate systems. The bounding boxes can also be used for different tasks such as: representing the segment as a probabilistic heat map [PJM+20] to account for the tumour infiltration [PAB+14, SCJ+12], prompting large/powerful models such as Segment Anything Model (SAM) [KMR+23], tracking algorithms [CWT+23].

3.5 Conclusion

What has been highlighted in this chapter is our thoughts on the challenges of segmenting brain tumours in 2D US images, with a preliminary study conducted to corroborate our hypothesis. For future work, larger curated and consensus annotated datasets of IOUS brain tumour images and volumes are needed to develop more accurate computer assisted boundary detection tools. This is likely to only be achieved through multi-site collaboration and pooling of data.

There are a few limitations of this study which we hope to be addressed in future work. Firstly, the number of annotated cases was small which is why we have withheld from identifying correlates - such as whether certain tumour types are more accurately segmented. Secondly, only single slices were used for segmentation as opposed to volumes. This is unlike the real-world use of IOUS where assessment of boundaries is based on live 3D sweeps of tumours and adjacent anatomy, plus often cross-correlation with preoperative MRI, both of which would help refine the accuracy of segmentations. However, we do comment that even though 3D information may be available, decisions are still biased by what is visible, which invariably would still be a 2D slice. Extending from this, there remains the recurring issue in neuro-oncology of there not being an accepted gold-standard ground truth for tumour boundaries and our use of integrated multimodal MRI and US to create the reference standard segmentations has to be an accepted compromise. Finally, whilst bounding boxes may serve as an efficient method to improve the detection of tumours, this is at the expense of specificity, which is important for the prevention of inadvertent removal of normal, functional brain tissue.

Chapter 4

Robust Fine-Tuning For Tumour Detection

4.1 Introduction

What amount of data is sufficient for neural network training? Although the requirement may vary, depending on factors such as the number of parameters and the specific task, the underlying question remains unresolved and perpetually unanswered. The theoretical answer is a dataset that covers the total distribution, which would facilitate neural network generalisability. Achievable for constrained or toys tasks. Unachieved in real-world tasks.

Ethical clinical translation can only be achieved when systems can behave and perform at the same level as their human counterparts. This can be boiled down to correct decision making, robustness to unknown, unexpected, or unlikely circumstances, and interpretability of the decision-making process. Current deep learning research for medical imaging has shown significant promise; however, tasks that are often solved have been simplified to accommodate the ease of data preparation and neural network development. For example, classifying histopathological image classification for bening vs. malignant [SOPH16]. Unlike what a histopathologist would do, which would include identifying cancer type, cancer grade, and characterisations. Segmentation of post-treatment MRI glioma [dVSG+24]. Where the type of tumour (e.g.,

glioblastoma, oligodendroglioma, astrocytoma), tumour grade, patient meta-data are ignored. In addition, MRI sequences are limited to T1, T2, T1Gd and FLAIR. Although these datasets are invaluable for benchmarking methods and developing toward clinical translation. They suffer fundamental flaws in replicating the real practical scenario, which include numerous more variables that may not easily be quantified or posed for deep learning. As such, deep learning successes tend to be limited to a small subset of examples relative to the whole set of clinical tasks. Here, a general assumption can be made that there is a correlation with the frequency of occurrence with likelihood of available data, and therefore progress in technical translation, in essence a long-tail issue. This inherently introduces significant perceptual biases in the translation community. Leading to unhelpful prophecies, most notably, "They should stop training radiologists now." Geoffrey Hinton 2017 [Muk17] - there is currently a global shortage of radiologists.

Data availability cannot be assumed as a guarantee. Many approaches have been developed to mitigate this limitation, including generative methods, meta-learning, and unsupervised/self-supervised. Examples include synthetic data generation to mitigate class imbalance and data scarcity for endoscopy multiclass segmentation [ZTAG24]. Segmentation of lesions was explored using the implicit model agnostic meta-learning optimisation framework under the few shot setting, which enables fast neural network training convergence from a small amount of data [KJH+21]. In [ZGCS23] a single image denoising diffusion model is proposed to synthesise lung ultrasound images. Where they demonstrate that this approach can be used to improve downstream tasks, specifically classification performance (healthy, pneumonia, COVID-19) by using the synthesised images for data augmentation. Most relevant to this work, [CBM+19] proposes self-supervised context restoration for pre-training model parameters on a larger unlabelled medical dataset, enabling improved fine-tuning on small datasets with limited annotations, under the assumption that initialising weights on data more similar to the target dataset is beneficial.

In this chapter, a study is constructed to explore a pressing but highly specialised application of ultrasound (US). That is, intraoperative ultrasound (IOUS) for brain tumour resection [DLGS+22]. This task has received relatively little attention, in part due to the lack of public

data [CESS+25]. Specifically, this study explores recent developments in robust fine-tuning [TDM+23] [TLSK23], which have mainly been aimed at maintaining large model performance on in- and out-distribution data after fine-tuning through constrained gradient descent. Following recent research direction [CESR+25], this study explores the specific task of object detection which has advantage over segmentation due to the minimisation of the impact of margin uncertainty [WDA+23a].

This study exclusively uses transformer-based architectures for object detection, specifically DETR [CMS+20] and DEF DETR [ZSL+20]. These approaches offer a distinct advantage over traditional architectures, such as the two-stage Faster R-CNN [RHGS15] or one-stage detectors like YOLO [RDGF15], by leveraging global context through the attention mechanisms. The global context is particularly critical for the detection of brain tumours, where margins often exhibit substantial aleatoric uncertainty, and tumours can vary significantly in scale and appearance. Furthermore, built-in attention layers can provide intuitive visualisations of model focus, constituting a truer form of explainability compared to techniques such as Grad-CAM [SDV+16], which only approximate activation. Finally, transformer-based detectors reduce the need for extensive hand-crafted feature tuning, further streamlining the detection pipeline for complex medical imaging tasks.

Furthermore, this study is interested in the impact of robust fine-tuning. Overcoming small data training complications is relatively simple with convolutional networks. However, attention-based architectures such as DETR often require more carefully constructed training frameworks [ZSL+20]. This issue, namely the performance gap induced by data efficiency between transformers and convolutional networks, has also been extensively explored in recent works [LXLZ22]. Consequently, examining methods to improve neural network training and outcome for DETR type architectures is essential. For a concise summary of the literature, the reader is referred to Section 2.2.2

A dataset containing 3705 images captured during brain tumour surgery was annotated to facilitate the experiments in this study. The results of this study show that, across both models, all folds and including all optimisation methods, an intersection over union (IOU) of

0.556 and F1 of 0.691 is achieved. Where specifically for the constrained gradient descent trained models exclusively, an IOU of 0.583 and F1 of 0.736 were recorded, highlighting both the feasibility of detecting tumours in ultrasound with good accuracy and the improvement capabilities when training with constraints on the parameter update process. Experiment extension was conducted to determine how the unconstrained and constrained models would handle unseen noise. In all experiments, the unconstrained models achieved 29 better IOU and F1 outcomes, while the constrained models achieved 67.

4.2 Method and Experimental Setup

4.2.1 Data

The data used for these experiments were captured during brain surgery at Charing Cross Hospital. Full, local, ethical approval for the collection of data was given by the HRA and Health and Care Research Wales (HCRW) authorities[1]. The dataset consisted of 15 videos, from 15 patients with various pathological conditions [WDA+23a], with 3705 frames extracted (20Hz) and annotated using bounding boxes - 3533 with tumour, 172 without. The annotation was performed by an experienced, but non-clinically trained, brain tumour IOUS reader. Reference annotations, a subset of frames from each video, were provided by a consultant neuroradiologist who also verified all final annotations. Example annotations are shown in Figure 4.1. What is highlighted specifically in these subfigures is the variability in the appearance of the tumours, both intra- and inter-scan / patient. This heterogeneity adds a unique level of complexity, unlike what is usually encountered in natural-scene tasks. Although the dataset is large in frame number and annotations, it is vital to highlight that fundamentally this dataset only comprises 15 patients / tumours, which therefore means there is limited tumour sampling, which cannot be representative of all brain tumours.

4.2.2 Robust Fine-Tuning

Standard gradient descent can be expressed as

$$\mathbf{W}_{t+1} = Opt(\mathbf{W}_t, \nabla L(\mathbf{W}_t)) \tag{4.1}$$

where $\mathbf{W}$ represents the trainable weight parameters of the neural network, t the previous iteration, and $\nabla L(\mathbf{W}_t)$ the parameter gradients given a loss function L.

[1]Study title - US-CNS: Multiparametric Advanced Ultrasound Imaging of the Central Nervous System Intraoperatively and Through Gaps in the Bone, IRAS project ID: 275556, Protocol number: 22CX7609, REC reference: 22/WA/0259, Sponsor: Research Governance and Integrity Team (RGIT).

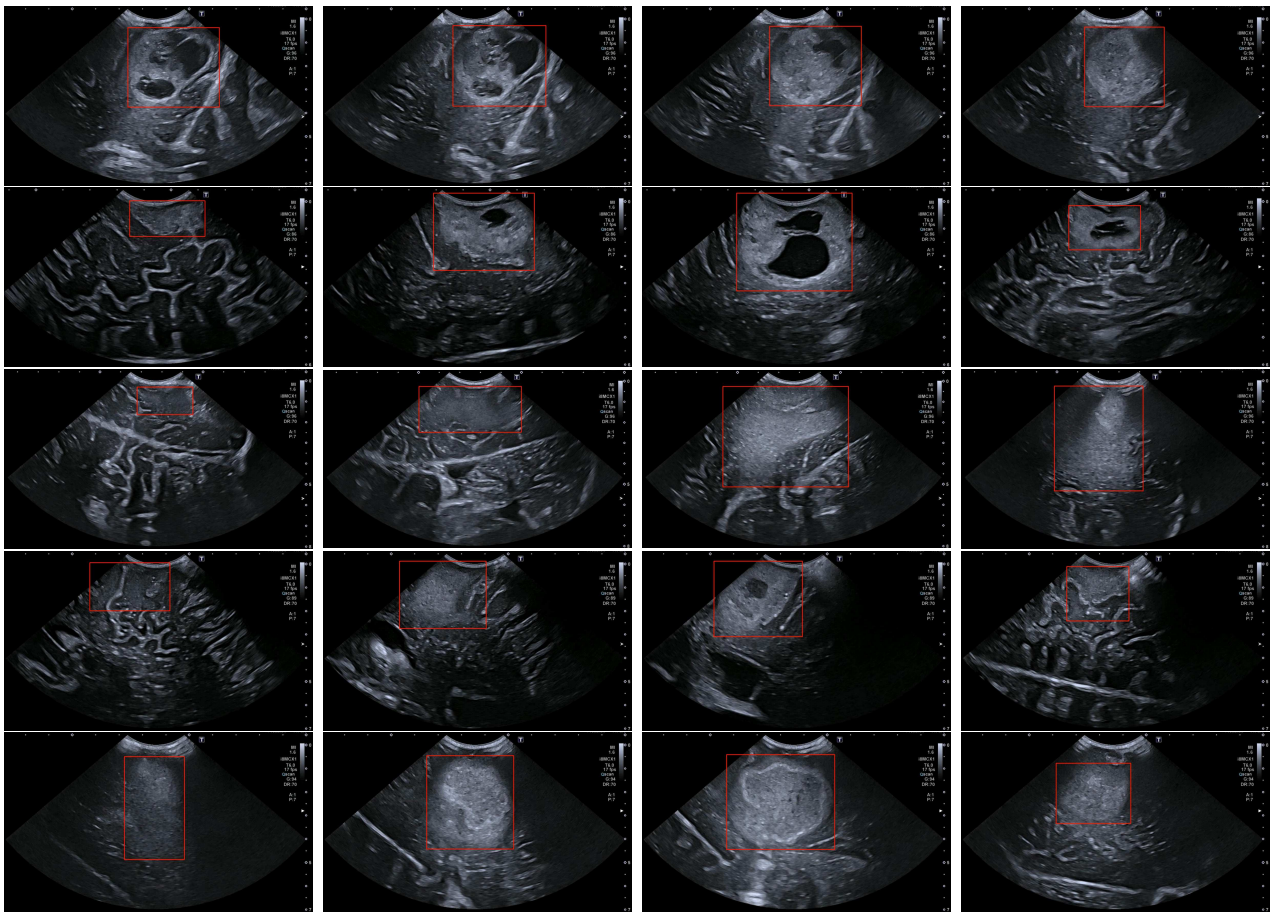


Figure 4.1: Sample ground truth annotations, where each row corresponds to a different scan (different patient) and going from left to right is going forward in time.

The method proposed in [TLSK23], which is a main focus of this chapter, introduces a gradient descent framework that remains anchored to the pre-trained weights through an adaptive and optimisable norm-ball constraint. This creates a dual optimisation framework that simultaneously minimises both the loss function gradients and the deviation of the updated parameters from the pre-trained weights. The original application of this method was to maintain performance in distribution and out of distribution. In this work, the application of this method is extended to explore the assumption that pre-trained models, which have been trained on large datasets, relevant or not to the fine-tuning task, have learnt a parameter set which is well structured, with inherently robust geometric properties, for image processing. Therefore, we alter the original narrative, and repurpose the method to explore it it is better to train a network for perturbation/deviation from these pre-trained parameters, than to allow the parameter set to change significantly; which will likely occur if fine-tuning is unconstrained.

$$minimise \, \|\nabla L(\mathbf{W}_t)\|, \|\mathbf{W}_{t+1} - \mathbf{W}_0\| \tag{4.2}$$

A summary description of the constrained gradient descent is presented.

$$\delta_{t+1} = Opt(\mathbf{W}_t, \nabla L(\mathbf{W}_t)) - \mathbf{W}_0 \tag{4.3}$$

where δ_{t+1} represents the current deviation of the unconstrained gradient descent update $Opt(..)$ from the pre-trained weights $\mathbf{W}_0$.

$$\nabla\gamma_{t+1} = \frac{k}{||\delta_t||_1} \sum (\nabla L(\mathbf{W}_t) \odot \delta_t) \tag{4.4}$$

$$\gamma_{t+1} = AdamOpt(\gamma_t, \nabla\gamma_{t+1}) \tag{4.5}$$

where $\nabla\gamma_{t+1}$ represents the gradient of the learnt projection constraint, calculated from the unconstrained gradients, and k is a gradient annealing factor. The scaling by the L1 norm of the deviation $\frac{k}{||\delta_t||_1}$ normalises the term, creating the norm ball constraint. The projection constraint γ_{t+1} is then updated using the Adam optimization step applied to γ_t.

$$\mathbf{W}_{t+1} = \mathbf{W}_0 + \frac{\gamma_{t+1} \cdot \delta_{t+1}}{||\delta_t||_1} \tag{4.6}$$

Lastly, this represents the adaptive norm-ball projection update, ensuring that the updated weights remain constrained relative to the pre-trained initialisation. The γ_t is initialised as $1e-9$. The terms δ and $AdamOpt$ are initialised as 0.

4.2.3 Training Framework and Comparisons Setup

All experiments were carried out using Python, Pytorch and Hugging Face, on a Linux computer with a single GeForce RTX 3080.

The dataset was used to create a 3-fold experimental setup, where the train/test split was performed at the patient level with a 10:5 ratio. Since each tumour and the corresponding video are distinct, this approach maximises data utilisation while maintaining a robust evaluation, despite the limited number of cases. The use of 3-fold cross-validation provides a thorough experimental setup and reduces the need for evaluation on additional datasets, which have limited public availability [CESS+25]. The images are of dimensions 1280x960.

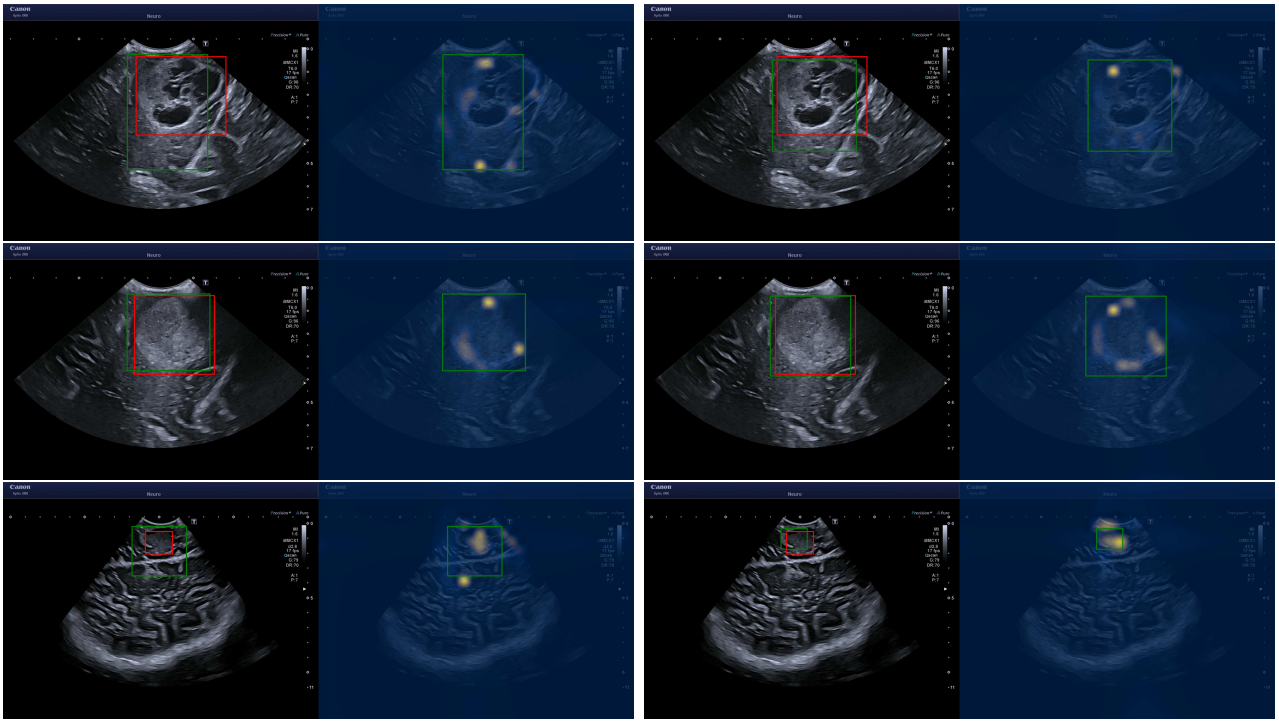
Both DETR and DEF DETR were pre-trained on COCO 2017 [LMB+14] with RESNET-50 [HZRS15] backbone, with the bipartite matching loss used for pre-training and fine-tuning [CMS+20]. All fine-tuning was performed for 20 epochs, where the batch size is set to 4, the number of queries to 20, and all experiments were fully seeded. As per the original paper, the robust fine tuning, when applied, is applied to all parameters except the head layers, which have to be reinitialised. The evaluation was performed using IOU and F1 metrics [PPD+21], where the F1 IOU threshold was set to 0.5. The best performing checkpoint on the corresponding test sets was taken, using the IOU score, for reporting. As there is only ever one object per image, and near all images contain a tumour, only the best output is evaluated, provided the prediction label is for tumour and not for background. Accordingly, the mean average precision is omitted due to the single-instance, single-prediction setup. This is determined by taking the bounding box with the corresponding highest output activation confidence. Due to its subsampling mechanism, DEF DETR inherently provides less informative attention maps. Consequently, this study focusses exclusively on DETR. The attention maps were constructed using the final cross-attention output, by averaging across the attention heads and reshaping and upsampling the attention sequence using bicubic interpolation.

Specifically for the experimental setup, ADAM and SGD [Rud16] are used, where their robust equivalents are denoted ADAM P and SGD P. For all optimisers, the learning rate is set to 2.0e-5, the weight decay to 5.0e-4. For SGD specifically, the Nestorov moment is used and set to 0.9.

Furthermore, no data augmentation was applied during fine-tuning. This enabled evaluation of the impact of noise on the test set performance, which can be perceived as a measure

of robustness. Although the augmentations used during this evaluation are known and are commonly used during training. Considering the diversity of noise and disturbances as well as variability in image capture and quality, in medical imaging. This approach of holding back these augmentations from training is done to control the validation, acting as an approximate of unknown real noise, disturbance, and variability (real-world unpredictability). Specifically, four sources of noise are chosen from the Albumentations library [BPK+18]: GaussNoise with settings std range = (0.2, 0.44), mean range = (0.0, 0.0), noise scale factor = 1, per channel = True; RandomBrightnessContrast with settings brightness limit = (-0.2, 0.2), contrast limit = (-0.2, 0.2), brightness by max = True, ensure safe range = False; MultiplicativeNoise with settings multiplier = (0.5, 1.5), per channel = False, elementwise = True; PixelDropout with settings dropout prob = 0.2, per channel = True, drop value = 0, mask drop value = None. The probability of applying the augmentation is set to 1 for all four augmentations.

4.3 Results



(a) Adam (b) Adam P

Figure 4.2: Example outputs from DETR on fold 3 test data. The red bounding box is the ground truth and the green is the best predicted. The brighter the yellow, the more attention.

4.3.1 Detection Performance

The results relevant to this subsection are Table 4.1 and Figure 4.2. The most significant result of this experiment is that the constrained gradient descent achieved better IOU and F1 results than the unconstrained networks, as shown in Table 4.1. To note, DETR fold 2 SGD failed to train, this can be contributed to the general instability when training using SGD. As all parameters were fixed, the metrics for this experiment are set to 0. Following the format unconstrained:constrained, the DETR IOU 0.475:0.577 and F1 0.571:0.716 and for DEF DETR IOU 0.581:0.590 and F1 0.720:0.756. Figure 4.3 highlights the impact of constrained gradient descent on the output models, where for all parameters, the constrained gradient descent produces parameters much closer to the pre-trained weights.

Table 4.1: Detection performance for the models produced using unconstrained and constrained gradient descent. The epoch number corresponds to the best performing checkpoint on the test set.

Fold	Optimizer	DETR			Def DETR		
		Epoch	IOU	F1	Epoch	IOU	F1
1	SGD	12	0.4923	0.5822	14	0.5394	0.5822
	SGD P	20	0.5820	0.7676	14	0.5084	0.5700
	ADAM	16	0.6165	0.7913	14	0.6382	0.8023
	ADAM P	12	0.6343	0.8073	16	0.6243	0.7517
2	SGD	–	N/A	N/A	14	0.5904	0.7472
	SGD P	18	0.5607	0.7758	14	0.6168	0.8428
	ADAM	20	0.6536	0.8300	20	0.5977	0.7998
	ADAM P	12	0.5019	0.5362	14	0.5591	0.7394
3	SGD	20	0.4378	0.4335	12	0.5476	0.6213
	SGD P	12	0.4819	0.5301	12	0.5721	0.7442
	ADAM	20	0.6508	0.7898	14	0.5711	0.7654
	ADAM P	20	0.7020	0.8810	18	0.6562	0.8858

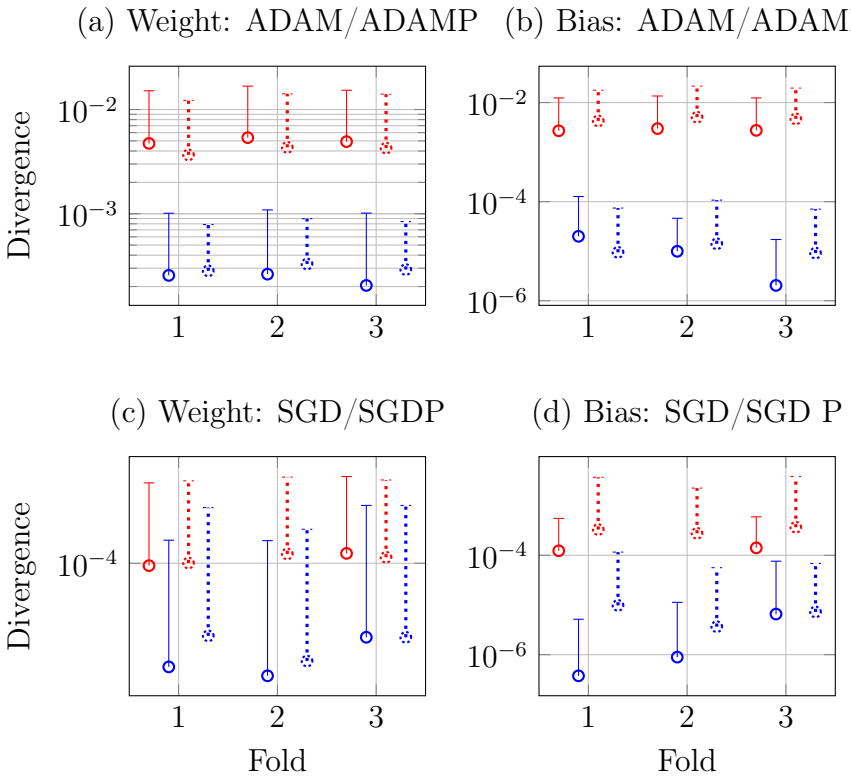


Figure 4.3: Impact of constrained gradient descent on the final model weights. The solid lines show DETR and the dotted lines DEF DETR. Unconstrained gradient descent is highlighted in red while constrained gradient descent with blue. The circle represents the mean difference between final model and pre-trained parameters, while the lines highlight the standard deviation. The y axis is log scaled.

Table 4.2: Gaussian Noise.

Fold	Opt	DETR		Def DETR	
		IOU	F1	IOU	F1
1	SGD	0.4919	0.5906	0.5114	0.5677
	SGD P	0.5901	0.7837	0.5026	0.5494
	ADAM	0.6118	0.7852	0.6090	0.7881
	ADAM P	0.6336	0.8035	0.6136	0.7205
2	SGD	N/A	N/A	0.5897	0.7294
	SGD P	0.5342	0.6397	0.6216	0.8418
	ADAM	0.6535	0.8310	0.6087	0.8096
	ADAM P	0.4662	0.4741	0.5797	0.7739
3	SGD	0.4286	0.4173	0.5481	0.6221
	SGD P	0.4812	0.5209	0.5327	0.6692
	ADAM	0.6466	0.7883	0.5741	0.7753
	ADAM P	0.6832	0.8547	0.6346	0.8935

Table 4.3: Brightness & Contrast.

DETR		Def DETR	
IOU	F1	IOU	F1
0.4546	0.5128	0.4507	0.4632
0.4966	0.5845	0.4824	0.5464
0.5810	0.7295	0.5625	0.7174
0.6075	0.7455	0.6095	0.7318
N/A	N/A	0.5131	0.5796
0.4925	0.5964	0.5286	0.6387
0.6298	0.8093	0.5517	0.6633
0.4743	0.4771	0.5161	0.6472
0.4057	0.3833	0.5309	.6051
0.4179	0.4227	0.4959	0.5842
0.5827	0.6685	0.5277	0.6751
0.6548	0.8233	0.5851	0.8022

Table 4.4: Speckle Noise.

Fold	Opt	DETR		Def DETR	
		IOU	F1	IOU	F1
1	SGD	0.3431	0.2053	0.4094	0.4121
	SGD P	0.4076	0.2701	0.4467	0.5120
	ADAM	0.6046	0.7867	0.5938	0.7935
	ADAM P	0.5936	0.6814	0.5308	0.6241
2	SGD	N/A	N/A	0.4333	0.4564
	SGD P	0.3926	0.4091	0.4989	0.5737
	ADAM	0.6145	0.8093	0.5754	0.7308
	ADAM P	0.4647	0.5086	0.5597	0.7416
3	SGD	0.3408	0.1963	0.4895	0.4930
	SGD P	0.4274	0.4173	0.4616	0.4298
	ADAM	0.6479	0.7944	0.5574	0.7608
	ADAM P	0.6749	0.8594	0.6299	0.8919

Table 4.5: Pixel Drop.

DETR		Def DETR	
IOU	F1	IOU	F1
0.4286	0.4838	0.4155	0.396
0.444	0.3586	0.4678	0.5479
0.5843	0.718	0.4019	0.4519
0.6118	0.7303	0.4812	0.5439
N/A	N/A	0.4665	0.485
0.3789	0.2888	0.5669	0.7284
0.587	0.7836	0.5143	0.5788
0.4928	0.5835	0.5448	0.7081
0.3953	0.3771	0.5559	0.6538
0.4258	0.398	0.4775	0.49
0.5635	0.6677	0.5526	0.7548
0.6382	0.7674	0.4140	0.6182

4.3.2 Inherent Robustness to Noise

The results of this experiment are highlighted in Tables 4.2, 4.3, 4.4, 4.5. Summing across all experiments, the unconstrained models outperformed the constrained 29 times, while 67 times vice versa. This clearly demonstrates the advantage of maintaining similarity to the pre-trained weights which would have inherent robustness due to the large original training dataset.

4.3.3 Attention maps

The example attention map outputs of the DETR ADAM/ADAM P fold 3 experiments are shown in Figure 4.2. The examples were chosen to highlight the impact of different textures, tumour sizes, necrosis, and definiteness of the tumour margin. Due to the lack of quantifiability of the behaviour of the attention maps, this evaluation is conducted qualitatively. What is most notable is the similarity between the output maps of the two fine-tuning methods. The attention mechanism for larger tumours will highlight the margin regions, while for smaller tumours it will highlight the entirety of the tumours.

4.4 Conclusion

This study presents an evaluation of fine-tuning approaches for transformer architectures for brain tumour detection in IOUS. Specifically, DETR and DEF DETR are trained using an unconstrained and constrained gradient descent methods. The US scans of 15 patients were annotated, producing a dataset of 3705 images. Standard object detection evaluation is conducted, and it was recorded that the constrained models outperformed their corresponding comparators. The evaluation was extended to robustness against unseen noise, to which again the constrained models achieved the better overall performance. Future work should focus on determining when enough data is sufficient to use unconstrained gradient descent, and specifically the limits of the benefit of constrained gradient descent. Limitations of the work are notably the evaluation on a single dataset and the use of a single architecture type. Future work should extend this set of evaluations.

Chapter 5

A Sparse Annotation Method Based On Confidence Grading

5.1 Introduction

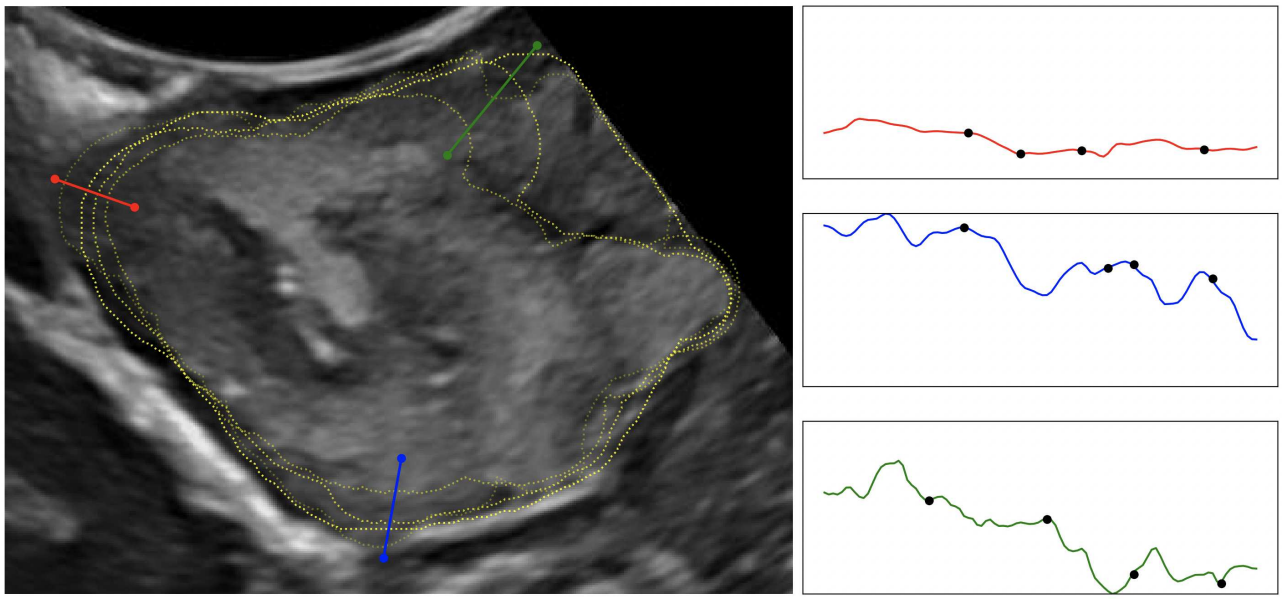


Figure 5.1: Tumour margin delineation uncertainty among four professional annotators. The three vectors show visible ambiguity over boundary point.

The annotation of objects of interest in medical images is invaluable from both a clinical and technical perspective. For normal anatomy or pathological tissue, the annotation of these

features can serve a multitude of purposes that can include helping to make treatment decisions (for example, long-term monitoring) [FNB+15], or postoperative evaluation (for example, extent of resection)[HFP+23]. In clinical practice, tumours can be annotated in many ways, such as diameter measurement [LWLF+22] or arrows for localisation [ST16]. The most precise form of annotation for differentiating pathological tissue from normal tissue is discrete segmentation [MJB+15]. This involves the binary separation of an object from its surroundings/background along the boundary of the object, or separation via different categories of the tumour and surrounding normal tissue such as enhancing, nonenhancing, oedema and necrosis.

In the clinical setting, ultrasound (US) is one of the most widely used imaging modalities. The NHS publication "Diagnostic Imaging Dataset Annual Statistical Release 2022/23" [Dix23] reports that between the years 2012/13 to 2022/23, US imaging accounted for $> 20\%$ of all recorded imaging activity. The second most commonly used modality behind X-rays. Where the common application of US was for the abdomen/pelvis.

The importance of US has resulted in published single- and multicentre data, which has been beneficial for both clinical education and technical research. In particular, motivated by recent successes with machine learning, annotated datasets have been curated for the task of semantic segmentation. Some examples of such are breast cancer [ADGKF20], cardiac structures [LSP+19], and foetus head [vdHdBdKG18]. A common denominator of these datasets is the clarity of the object of interest, where the boundaries of the objects are well defined with limited visual uncertainty.

Intraoperative ultrasound (IOUS) has become a commonly used tool to assist in surgical planning, detection and localisation of pathological tissue, and margin delineation [LGK+21]. Recent evidence suggests that IOUS might be particularly useful in improving the likelihood of maximal safe resection during brain tumour surgery [DLGS+22]. Although a useful tool, due to the limited standardisation of training and the difficulty of interpreting images due to the confined field of view, the annotation of brain tumours in IOUS images can be difficult and result in variance between observers [WDA+23a]. This is further compounded in the case of diffuse tumours, where there is often an extremely poor separability of the tumour from the normal

surrounding tissue due to the graduated infiltrative nature of the disease and the inherent lack of a boundary.

This chapter explores the use of probabilistic annotation of brain tumours as an alternative to discrete annotation. Ill-definition, diffusion, infiltration of a tumour and its margin ultimately results in an incompatibility with discrete annotation. What is argued in this chapter is that annotation should be performed, not on a pixel level, but rather spatially and locally. The key contributions are as follows.

- A confidence-based annotation protocol and method is proposed.
- Training tumour segmentation networks with clinically relevant soft labels improves neural network calibration, compared to training with hard labels.

5.2 Methodology

5.2.1 Annotation Protocol

For the construction of the confidence map of the object, the annotator can use three criteria of scatter points - high, medium, and low confidence. Each scatter point is used to delimit the edge of the corresponding level of confidence over which a heat map is to be constructed. Based on computer vision and radiological theory, a protocol for defining these points has been constructed. Inspiration has been taken from radiological annotation approaches that differentiate the different layers of the tumour and the relevant surrounding tissue, as was done in [MJB+15]. Principally, this protocol encourages sensitivity over specificity due to possible extreme levels of aleatoric uncertainty along the tumour margins.

Areas of high confidence can be perceived as where the tumour mass is well defined with expected low levels of interobserver annotation variance - low aleatoric uncertainty in the pixel-level features. Where the granularity of US is sufficient to make the decision on whether to resect. This can be contributed to the area belonging to the hypercellular components of the tumour, which is usually hyperechogenic and thus considerably brighter than it's surroundings. Furthermore, where the texture is anomalous, significantly homogeneous or heterogeneous relative to the surrounding normal tissue. Or, where edges are well defined by calcification or necrosis.

The medium confidence annotation should primarily be associated with the edge of the fuzzy margin. Here there is likely a large degree of interobserver variability, and the margin aleatoric uncertainty is already high. For these zones, the decision to resect is difficult to make using purely the US image as a reference, where granularity begins to become or is unreliable. In particular, for a diffuse tumour, this is somewhere along the gradient between a clear tumour core and perceived normal tissue. For a well-defined tumour, this can be areas around the clear margin of the tumour that have the continued visual properties of the tumour. For example, continued brightness and or continued anomalous texture. This can also be due to obscuration or acoustic shadowing. This level of confidence is likely the hardest to annotate.

The distribution of low-confidence annotations will occur sparsely outside of the clear tumour mass and the "clear fuzzy margin". The margin aleatoric uncertainty and interobserver annotation variance are significant. The US granularity is not sufficient to make actionable decisions on resection. These areas will likely reflect the surrounding abnormal tissue that is a variable combination of oedema and less cellular tumour infiltration that correlates with the perilesional, non-enhancing abnormal signal observed on MRI [WRR+22]. Pragmatically, these would be the areas that would not typically be resected, as it is thought to reflect predominantly oedema, but where there remains an uncertainty and risk of residual tumour cells.

5.2.2 Confidence Map Construction

Once the annotation is complete, the following sets of coordinates are created: high likelihood $\mathcal{C}_{\mathcal{H}} = \{(x_a, y_a)\}_{a=1}^{n_H}$, medium likelihood $\mathcal{C}_{\mathcal{M}} = \{(x_b, y_b)\}_{b=1}^{n_M}$, and low likelihood $\mathcal{C}_{\mathcal{L}} = \{(x_c, y_c)\}_{c=1}^{n_L}$.

The mesh grid $\mathcal{G} = \left\{(x, y) \mid x \in \left[0, \frac{W}{\lambda}, \frac{2W}{\lambda}, \ldots, W\right], y \in \left[0, \frac{H}{\lambda}, \frac{2H}{\lambda}, \ldots, H\right]\right\}$ is then defined, given the width W and height H of the image, subsampled using λ. For this work $\lambda = 100$ provided a sufficiently smooth map.

The mesh grid is then weighted by aligning annotation coordinates with grid points.

$$w(x, y) = \begin{cases} \beta_H & \text{if } (x, y) \in \mathcal{DT}(\mathcal{C}_{\mathcal{H}}), \\ \beta_M & \text{else if } (x, y) \in \mathcal{DT}(\mathcal{C}_{\mathcal{M}}), \\ \beta_L & \text{else if } (x, y) \in \mathcal{DT}(\mathcal{C}_{\mathcal{L}}), \\ 0 & \text{otherwise.} \end{cases} \tag{5.1}$$

Where Delaunay triangulation $\mathcal{DT}$ is used to determine which grid points correspond to each likelihood region. $\beta_H, \beta_M, \beta_L$ was assigned values $1, 0.5, 0.1$.

The confidence map $\mathcal{CM}$ is finally constructed by applying a Gaussian Kernel Density Estimation (KDE), with normalisation, over the mesh grid $\mathcal{G}$ with a minimal bandwidth factor

$\sigma = 0.1$ chosen to prevent the over-influence of high confidence and further reduction of spatial specificity.

$$f(x,y) = \frac{1}{n\sigma^2 2\pi} \sum_{i=1}^{n} w(x_i, y_i) e^{-\frac{1}{2}(\frac{(x-x_i)^2+(y-y_i)^2}{\sigma^2})} \tag{5.2}$$

$$\mathcal{CM} = \frac{f(x,y) - \min(f)}{\max(f) - \min(f)}, \tag{5.3}$$

where n is the number of points in $\mathcal{G}$. Overall, the set of tunable parameters is $[\lambda, \beta_H, \beta_M, \beta_L, \sigma]$, which corresponds to [meshgrid size, annotation class weighting, KDE bandwidth].

5.3 Experiments And Results

This section will describe the evaluation setup that has been used to validate the usability of this method. Through, the implicit prediction of interoberserver variance as well as down stream data learning application. Sample annotations are displayed in Figure. 5.2.

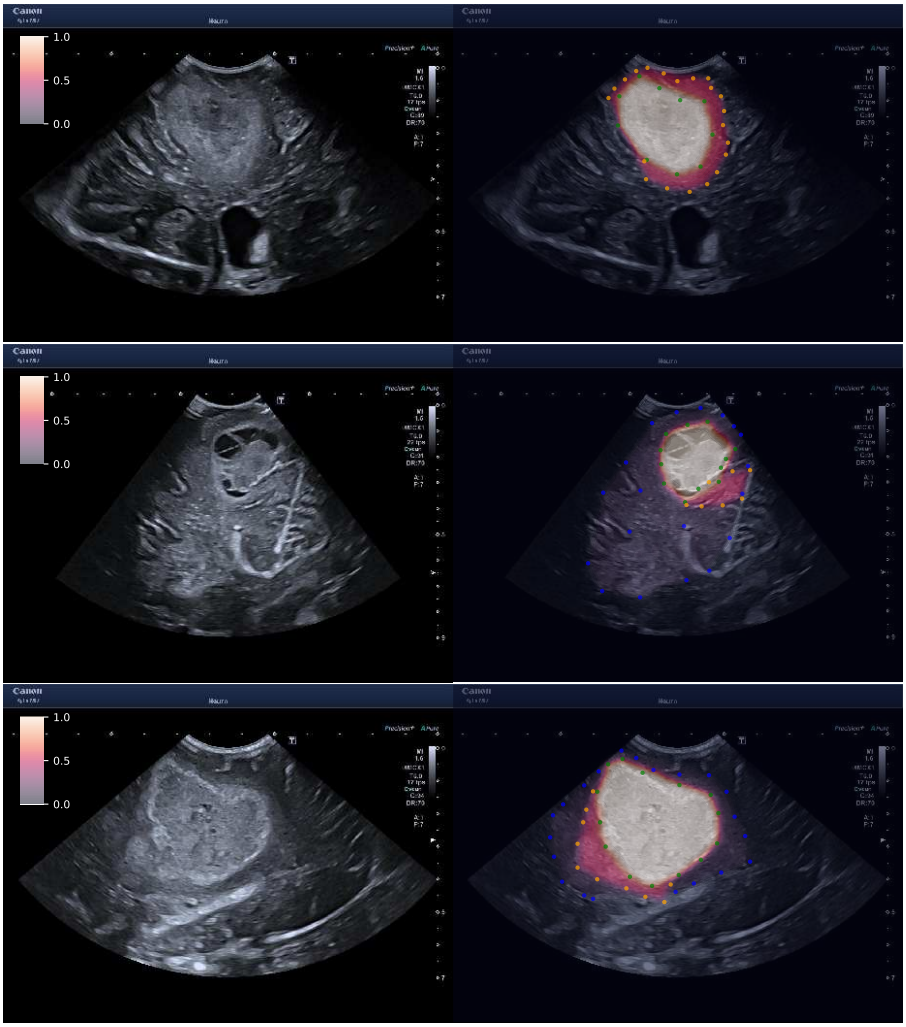


Figure 5.2: Sample ground truth annotations.

5.3.1 Confidence Map And Interobserver Variance Similarity

Data from [WDA+23a] has been used in this study to measure the correlation of confidence annotation with interobserver variance. The confidence annotations are performed with reference to the neuroradiologists annotations, which is defined in the cited study as their ground truth due to annotation in conjunction with multiple modalities, with the annotations by the

neurosurgeons unseen. The neuroradiologist annotations are used primarily for localisation, with the annotations strictly adhering to the defined protocols. In particular, this means that not all neuroradiologist annotations are described as high confidence.

5.3.1.1 Metric Similarity

To start, a basic image-wise comparison is performed between the confidence maps and the interobserver variance, denoted $Agreement = \frac{A_1+A_2+A_3+A_4}{4}$, where A is an individual binary annotation. Two metrics are used, the Brier score and the Soft Dice score, both produce scores lie in range $[0, 1]$, where lower values indicate better performance. Using both metrics accounts for the pixel calibration (Brier score is approximately equivalent to mean squared error but for probabilistic prediction accuracy) and the segment similarity (Soft Dice score is the variant of Dice score for soft/fuzzy annotation overlap analysis).

$$\text{Brier} = \frac{1}{n}\sum_{i=1}^{n}(p_i - y_i)^2 \tag{5.4}$$

$$\text{Soft Dice} = 1 - \frac{2\sum_{i=1}^{n} p_i y_i}{\sum_{i=1}^{n} p_i^2 + \sum_{i=1}^{n} y_i^2} \tag{5.5}$$

where n is the total number of pixels to be evaluated across all images. Where p is the prediction (confidence map) and y the ground truth label (agreement map), both in the range $[0, 1]$. The Brier score is calculated for both the entire image (considering foreground and background) and exclusively for the foreground (to avoid class imbalance), where the foreground is defined as the union of the agreement and confidence map. The Brier score with background is 0.00762 and without 0.0437. The Soft Dice is calculated on the entire image, and an average score of 0.076 is recorded. Both metrics indicate a strong agreement between the confidence map and *Agreement* maps.

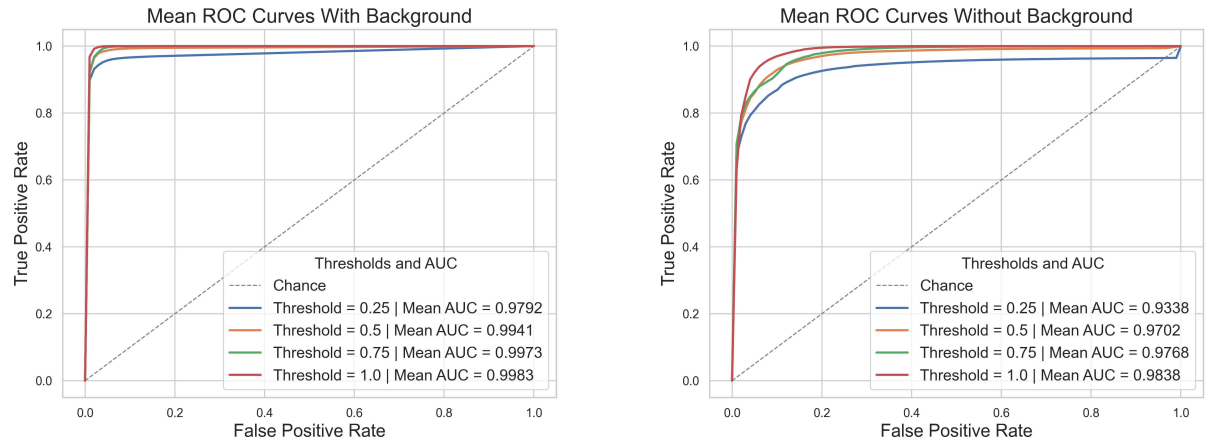


Figure 5.3: Mean Receiver Operator Characteristic (ROC) curves across threshold interobserver agreement levels. Left subfigure shows the ROC will all pixels included, while the right subfigure shows the ROC only using union area between the two annotations.

5.3.1.2 Similarity Across Threshold Levels

To further assess how well the confidence maps align with the *Agreement* maps. The Receiver Operating Characteristic (ROC) curve are computed at four different binary thresholds of the ground truth, $[0.25, 0.50, 0.75, 1.00]$. The ROC is calculated for both the entire image and exclusively for the foreground. The results of this ROC analysis are shown in Figure. 5.3. For both cases, the Area Under the Curve (AUC) is greater than 0.9, with the AUC improving as the threshold increases, which can be attributed to the reduced margin uncertainty. This result adds further evidence that the proposed protocol for creating confidence maps is effective in estimating interobserver variance.

Further threshold-based analysis is conducted, where in this instance, both the *Agreement* and corresponding confidence map are jointly threshold binarised. This evaluation differs from the previous one by evaluating the similarity between the binarised confidence map and the binarised *Agreement* map across multiple confidence levels. This joint thresholding approach effectively measures how well the confidence maps (and the annotation protocol) are calibrated to reflect the variability in interobserver agreement rather than just evaluating the overlap between annotations. The four evaluation metrics for this are Precision, Recall (TPR), F1-Score and Accuracy (ACC). The principal metrics that can be derived from a binary confusion matrix.

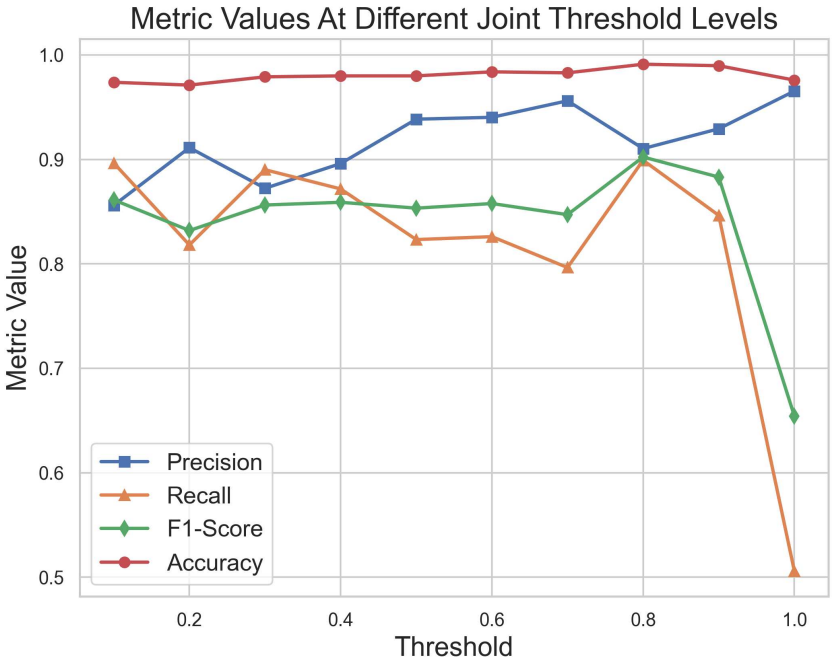


Figure 5.4: Metric performance across different threshold levels, where the thresholds are applied to both annotations.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{5.6}$$

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5.7}$$

$$\text{F1-Score} = 2\frac{\text{Precision} \cdot \text{TPR}}{\text{Precision} + \text{TPR}} \tag{5.8}$$

$$\text{ACC} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{5.9}$$

Where TP is True Positives, TN is True Negatives, FP is False Positives, and FN is False Negatives. The results of this experiment are shown in Figure. 5.4, where is is notable that all metrics highlight good similarity except at the 1.0 threshold. Where there is reduced performance in TPR and F1-score. This can be attributed with an increased number of false negatives, due to

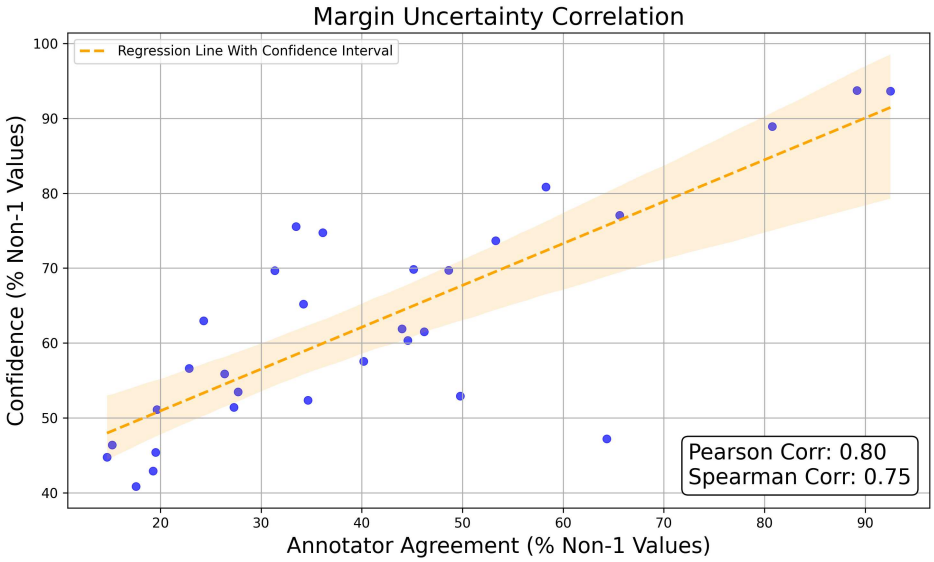


Figure 5.5: The relationship between annotator variance and confidence annotation at the tumour margin (non-1 values) shows a clear linear trend, supported by correlating statistical evidence.

the KDE's smoothness and therefore a reduced number of ones.

5.3.2 Margin Uncertainty Correlation

Extending the previous evaluation based on binarisation and segment analysis. Analysis of the percentage of non-1 values in both confidence and interoberserver variance is conducted, focused on the margin uncertainty correlation between confidence map and *Agreement* map.

$$\psi_{\text{conf}} = \frac{|\{(x,y) \in I \mid 0 < \text{conf}(x,y) < 1\}|}{|\{(x,y) \in I \mid 0 < \text{conf}(x,y)\}|} \times 100$$
$$\psi_{\text{agree}} = \frac{|\{(x,y) \in I \mid 0 < A(x,y) < 1\}|}{|\{(x,y) \in I \mid 0 < A(x,y)\}|} \times 100$$

where I denotes an image, conf the confidence map and A the agreement map. The metrics for this analysis are the Pearson correlation coefficient r and Spearman's rank correlation coefficient ρ; respectively, this measures the linearity and monotonicity of the relationship of margin uncertainty in both sets of annotations.

$$r = \frac{\sum_{i=1}^{n} \left(\psi_{\text{agree},i} - \overline{\psi_{\text{agree}}}\right)\left(\psi_{\text{conf},i} - \overline{\psi_{\text{conf}}}\right)}{\sqrt{\sum_{i=1}^{n} \left(\psi_{\text{agree},i} - \overline{\psi_{\text{agree}}}\right)^2}\sqrt{\sum_{i=1}^{n} \left(\psi_{\text{conf},i} - \overline{\psi_{\text{conf}}}\right)^2}} \tag{5.10}$$

$$\rho_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \tag{5.11}$$

where d denotes the rank difference between the two margin uncertainty observations; visualised with the blue dots in the results plot in Figure 5.5, between both annotation sets. Where the measured Pearson correlation was 0.80 and the Spearman correlation 0.75. The scatter plot shows the correlation between confidence map and interobserver variance, with a linear regression model fitted to the data with a 95% confidence interval. The conclusion of this experiment is that the proposed annotation method will correctly increase the spread of confidence as the size of the tumour and its margins (and uncertainty) increase spatially.

What can be concluded from this set of evaluations is that there is a correlation between confidence maps and unseen interobserver variance. Not only does this affirm the benefit of this protocol and method of annotation in minimising the annotator-level epistemic uncertainty through treating the annotation as a local density spatial task rather than at a discrete pixel/voxel level. But also, these results suggest that this annotation strategy would greatly reduce time taken, as there is a reduce or even no requirement for multiple annotators.

5.3.3 Downstream Application: Soft Label Learning

The proposed annotation method is faster and simpler than traditional segmentation and can facilitate dataset generation. Confidence annotations, which cover a larger pixel area and incorporate likelihood grading, may provide more nuanced information about tumours. In deep learning segmentation tasks, probabilistic pixel-wise classification is typically performed using sigmoid or softmax activation functions, with discrete annotations.

Studies, such as [SVI+15, MKH19, VSV20] have explored the benefit of using soft labels, continuous labels in the binary case, and non-one-hot vectors in the multiclass, for regularisation.

Where [GLCA20] found that soft labels improved the sensitivity of segmentation networks, particularly for small objects in medical images.

Experiments are performed by training a standard UNet model on both the confidence (soft) and discrete (hard) annotations, where the latter was the annotation from the neuroradiologist. Individually trained to assess the downstream benefit of this method for annotation.

For this experiment, a UNet model with 4 encoder/decoder layers and a maximum channel depth of 64 was trained using sigmoid activation in the final layer. Training was performed for each annotation individually, within a 5-fold cross-validation framework. With training conducted using several standard loss functions; Brier and Soft Dice (Equation. 5.4, 5.5), along with the following, to ensure that the evaluation is not biased towards a specific loss:

$$\text{Binary Cross Entropy} = -\frac{1}{n}\sum_{i=1}^{n}\left[y_i \log(p_i) + (1-y_i)\log(1-p_i)\right] \tag{5.12}$$

$$\text{Huber} = \frac{1}{n}\sum_{i=1}^{n}\begin{cases} 0.5\cdot(p_i-y_i)^2, & \text{if } |p_i-y_i| < 1 \\ |p_i-y_i| - 0.5, & \text{if } |p_i-y_i| \geq 1 \end{cases} \tag{5.13}$$

$$\text{Tversky} = 1 - \frac{\sum_{i=1}^{n} p_i y_i}{\sum_{i=1}^{n} p_i y_i + \alpha\sum_{i=1}^{n} p_i(1-y_i) + \beta\sum_{i=1}^{n} y_i(1-p_i)} \tag{5.14}$$

where the Tversky parameters are$\alpha = 0.3, \beta = 0.7$ [SEG17]. The metrics used for comparison are the same as in Section. 5.3.1.1. Where Brier score is particularly informative of calibration calibration performance [Bri50]. Results of this are shown in Table. 5.1. Displayed in Figure. 5.6 is a sample of outputs for the Binary Cross Entropy test set. Qualitatively, the soft trained network produces smoother confidence. Notably, more likely to activate with low confidence in high uncertainty regions. Unlike the hard trained network which tends to behave more discretely. Notably, the necrotic sphere in the bottom row is ignored by the hard trained network but low confidence is activated with the soft trained network. Which is a more desirable when considering that the necrotic sphere, in both networks, is surrounded by activated pixels.

Table 5.1: Training a UNet on soft and hard annotations with varying loss functions. The results correspond to the average across the 5-folds. For all loss functions, the soft trained networks achieves better test performance than the hard trained networks.

Loss Function	Brier		Soft Dice	
	Soft	Hard	Soft	Hard
Brier	0.0179	0.0234	0.2219	0.2989
Soft Dice	0.0247	0.0270	0.2236	0.2868
Binary Cross Entropy	0.0176	0.0220	0.2132	0.2833
Huber	0.0180	0.0224	0.2269	0.3016
Tversky	0.0247	0.0285	0.2131	0.2917

5.4 Discussion

Fully informed diagnostic decision making with MRI requires multiple types of sequences. As stated in [PG11], reliance on individual sequences such as T_1-weighted or T_2-weighted, will underestimate the spread of cancer in the surrounding areas of the determined region of interest. For example, [SBD+10] measured that the use of proton magnetic resonance spectroscopic imaging, on average, identified tumour margin extensions greater than 20% from what was identifiable using T_2 weighted images. We extrapolate this with the use of standard B-mode US. An example of this can be seen in the use of superb microvascular imaging (SMI). In [CXW+24], it was seen that the use of SMI, particularly for high-grade gliomas, improved the visualisation of the tumour margin and was associated with an increased likelihood of gross total resection. This clinical perspective further emphasises the limitation of B-mode segmentation and the issue of discretely determining the edge point. The evidence of margin difference when using multiple modalities or sequences, indicates not only that discrete margin delineation is not achievable purely using B-mode imaging. However, even with multiple sources of information, the margin uncertainty is only reduced, not eliminated, because of the infiltrative properties.

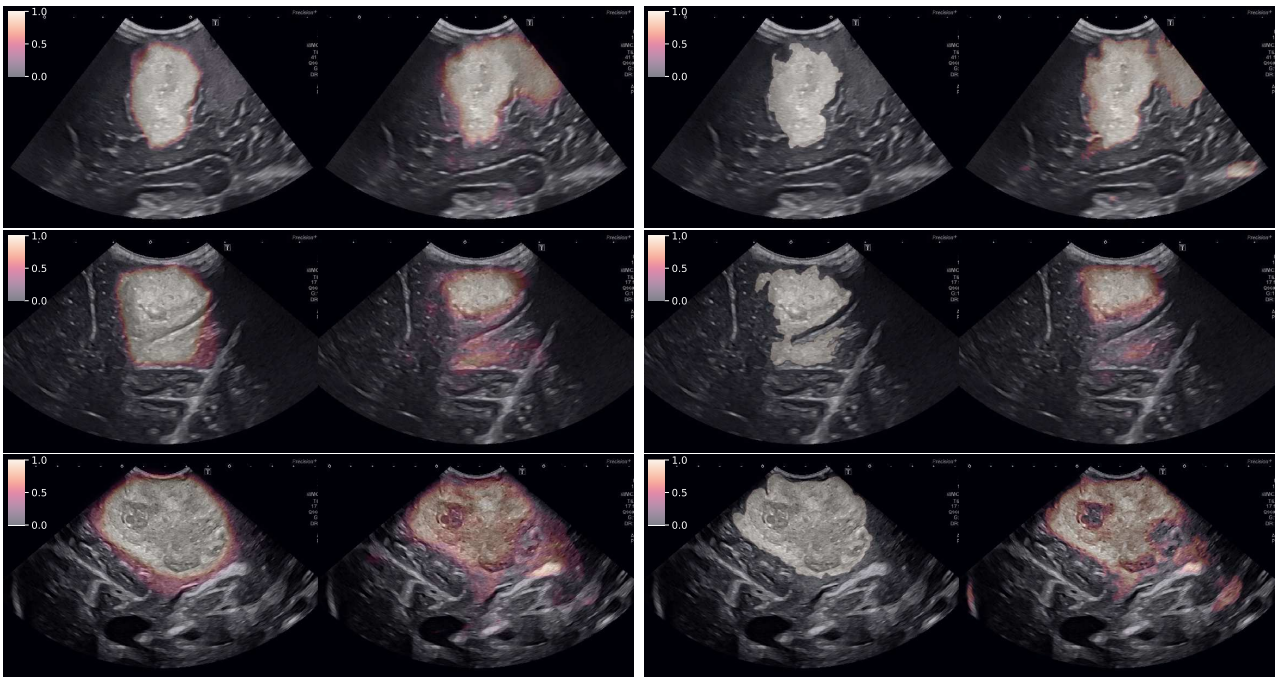


Figure 5.6: Test set output comparison from training on soft (left column) and hard (right column) ground truth. Where in each column, the left image is the ground truth and right the UNet output.

5.5 Conclusion

The challenge of discretely annotating diffuse, infiltrative, brain tumours in B-mode US is demonstrated. A method for confidence-based annotation is proposed with a protocol derived from computer vision and radiology theory. The method enables sparse and fast annotation which should facilitate rapid dataset generation. The confidence based protocol reduces and/or removes the need for averaging interobserver annotations - through minimising the annotator level epistemic uncertainty by the design of the annotation protocol and method. Evaluation is performed to measure how well the protocol and KDE process can, implicitly, estimate the interobserver variance of 4 professionals, discretely annotating the corresponding images. A reliability diagram is created and an average bin accuracy of 0.1 was measured which indicates correlation between the confidence method and the interobserver variance. In particular, a linear relationship was measured along the tumour margin. With a Pearson correlation of 0.8 and Spearman correlation of 0.75. The downstream application of the confidence annotations is investigated. A UNet is trained for segmentation on both the neuroradiologist discrete annotations (hard label) and the confidence annotations (soft label). Across all evaluation

folds, the soft labels produced a better calibrated network with better Brier scores.

The limitations of this study include the potential bias in the analysis of the relationship between confidence annotation and interobserver variation. Although neurosurgeon annotations were not seen during the confidence annotation process, the same authors participated in the prior study, which may introduce bias due to the influence of memory. However, this bias was controlled through strict adherence to the annotation protocol and by anchoring the localisation process on top of the neuroradiologist's annotations.

Future work to remove interposed normal tissue from the confidence map could be explored, perhaps using sink points.

Part II

Probe Manipulation Analysis

Preamble

Due to the ease and diversity in clinical application of ultrasound, the exploration of robotic development for fully automated and assisted scanning has shown growing interest within the research community and also promising commercial viability. However, the vast majority of published work has concentrated on non-invasive applications, resulting in a somewhat void in the space of IOUS, probably because non-invasive applications are easier to create experimental and evaluation setups for.

This PhD was carried out as one half of the Imperial - TUM Joint Academy of Doctoral Studies (JADS), "Confidence-Driven Robotic Ultrasound Tissue Scanning for Surgical Resection Guidance" project. My primary responsibility was the development of methods to facilitate robotic scanning through image guidance. What is unique about this robotic application is that the scan is performed on intraoperative tissue, notably brain tissue, which presents significant and unique challenges compared to non-invasive US scanning. The tissue is extremely delicate, where absolutely minimal force is required. In addition, the space allowed for the scan, dictated by the craniotomy, is highly variable per surgery.

Replicating the performance of our consultant neuroradiologist. The robotic system has been designed to use purely US feedback to guide the robotic scan. This has meant that - because of the safety critical nature of the task - the method had to be designed to be robust, interpretable, and real-time. Due to the complete lack of available data and the issues with interpretability of black-box methods, this ruled out any data learning approach. As previous methods in the literature were deemed inappropriate, I chose to design a completely new method, to identify the soft boundaries of visible brain tissue (such as brain convolutions and sulci, etc.).

This Part outlines this novel method, in which I perform an extensive evaluation of both the performance and robustness of the methodology, as well as the application for robotic US scanning. This resulted in the first major milestone of the JADS collaboration. In the context of this thesis, one of the first robotic solutions for the automation of US brain scanning is presented, which could be used in a fully automated or a human-in-the-loop system. This

can be applied to support neurosurgeons, both in and out of surgery, who may lack sufficient training or experience. Furthermore, when a neurosurgical team is overstretched, for example, fatigued due to a long operation or hospital understaffing. Future work will be published, led primarily by the team in Germany, which expands the experiments using my proposed method into a realistic intraoperative environment using porcine cadavers. In addition to proposing and demonstrating a novel and sophisticated control methodology.

Chapter 6

A Method For Evaluating Ultrasound Probe-Tissue Contact During Brain Surgery

6.1 Introduction

Intraoperative imaging is an important tool for surgeons in decision-making during surgery. The low cost of ultrasound (US) machines and the ease of integration into surgical settings has led to enthusiasm for intraoperative ultrasound (IOUS) [SBP+17].

One particularly challenging aspect is balancing optimal probe-tissue contact and orientation, without causing inadvertent pressure-related damage. Contact related acoustic shadow (AS) can occur due to interposed gas or contact breaking, which reduces image reliability through tissue obscuration. This is particularly challenging in the case of brain surgery [DLGS+22]. To demonstrate the importance of good contact for US image usability, two tumour segmentation images are presented in Fig.6.1, one that is affected by shadow and one that is not affected. The top row shows an example of suboptimal US imaging of a brain tumour, where there is incorrect focus and obscuration of the left side of the tumour from acoustic shadowing (highlighted with

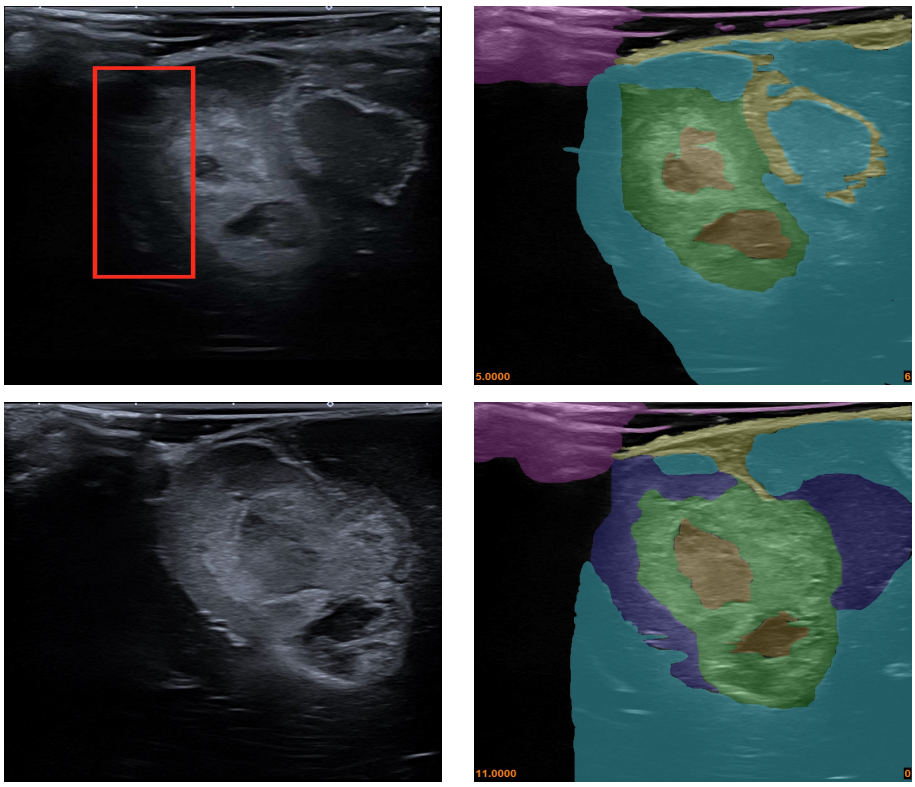


Figure 6.1: Ultrasound segmentation of IOUS images. The top row shows suboptimal US image capture, and the bottom row shows good US image capture. Annotations: green = solid tumour, orange = tumour necrotic core, yellow = cerebral sulci, dark blue = oedema, turquoise = surrounding normal brain, yellow = coupling fluid/probe brain interface.

a red box). Acoustic shadowing results in incomplete visualisation of the tumour, causing inaccurate segmentation with an artifactual straight-line border along the left margin of the lesion. The bottom row, in comparison, shows a well-focused, complete visualisation of the same brain tumour with clear delineation of the tumour boundaries, which greatly facilitates accurate segmentation.

Recent advances in IOUS have raised interest in algorithmic analysis. For example, detection of bone shadow [BCML16, APH20] calcification [dSFSY+06, LKHK18], and general shadowing [HCMC10, MHM+19, MASC22]. The main challenge with these methods is that they do not examine probe-tissue coupling, Figure 6.2 presents an example of where shadowing does not always equal poor contact. To address this limitation, the algorithm in [KWKN12a, CKN17, JGZ+20] produces a confidence map of signal attenuation, using a random walk-based method. The limiting factor for the latter method is the assumption that hyperechoic lines correlate with reduced signal attenuation. However, this is not a well-generalised assumption. In this

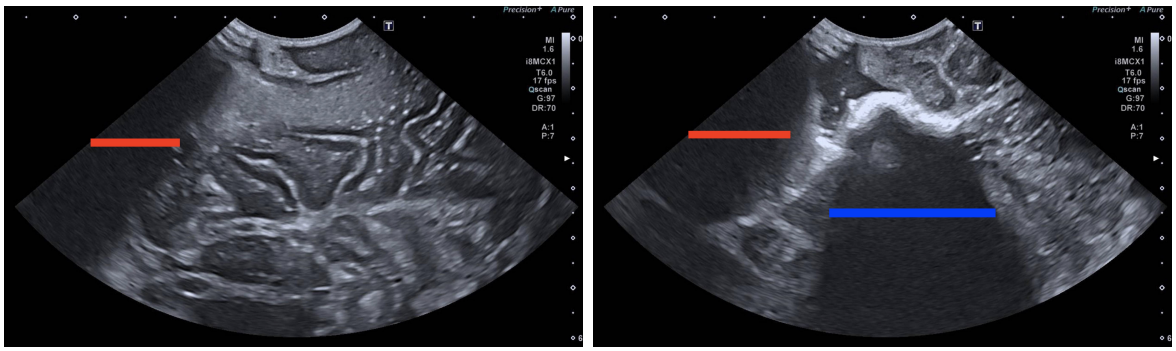


Figure 6.2: Sample annotations of b-mode ultrasound images. The red line delineates acoustic shadowing secondary to poor contact whilst the blue line defines acoustic shadowing related to other causes (in this case echogenic blood clot in a resection cavity).

work, we propose a method for identifying and topologically representing visible tissue in US for the detection of contact related AS and confidence in perceptual salience. Our contributions are as follows:

1. A method for identifying visible tissue within US images is proposed, using iterative filters and topological representation, that can be used for probe-tissue contact-related AS detection.

2. A novel framework is proposed for the analysis of the confidence estimate of ultrasound probe-tissue contact.

6.2 Methodology

6.2.1 Detection of tissue features

To detect areas of high spatial intensity variation, linear US images $I \in \mathbb{R}^{H \times W}$ are iteratively processed using Gaussian filters, with a vertical bias, following:

$$\begin{gathered} G_f = h_1 * G_{f-1}, \\ h_1 = (1/\sqrt{2\pi\sigma_u} e^{-u^2/2\sigma_u^2})(1/\sqrt{2\pi\sigma_v} e^{-v^2/2\sigma_v^2}), \\ \sigma_u >> \sigma_v, G_1 = I \end{gathered} \tag{6.1}$$

where, $G \in \mathbb{R}^{H \times W}$ denotes an output of the Gaussian filter, f the iteration set, and σ the standard deviation of the Gaussian filter stack. The image coordinates are u, v, where v increases with signal depth along scan lines (image columns), perpendicular to the transducer at the top. The Gaussian smoothing is biased vertically, as anatomical changes are more likely to be exposed in the vertical direction. The function is linear and the iterative processing is associative $(G * h_1) * h_1 = G * (h_1 * h_1)$ and therefore equal to a single convolution with a wider standard deviation $\hat{\sigma_u} \approx \sqrt{f}\sigma_u$. Unlike direct edge detection methods (Laplacian or Sobel), which are not robust to noise and do not offer a smooth response. Each output G_f of the Gaussian filter sequence is stacked and the standard deviation is taken for each pixel across the stack following:

$$S_{u,v} = \hat{G} * \sqrt{\frac{\sum_{f=1}^{len(f)+1}(G_{f,u,v} - \frac{\sum_{f=1}^{len(f)+1} G_{f,u,v}}{len(f)+1})^2}{len(f)+1}} \tag{6.2}$$

where, $S \in \mathbb{R}^{H \times W}$ is the output of the standard deviation. $\hat{G}$ is an extra smoothing filter for the standard deviation output with $\sigma = 1$, to remove outliers.

Points with high spatial variation correspond to areas that have been most affected by the smoothing process. Hence, these points are separated by thresholding the standard deviation output following:

$$\chi = \{(u, v) \mid S_{u,v} \geq \gamma\}, \gamma = round(\sigma_u \times \sqrt{len(f)}) \tag{6.3}$$

the γ formulation is elaborated on in Sec. 6.4.2.2. Points of high spatial variation are represented as a 2D point cloud χ.

To reduce the computational load for building the topology and to remove outliers, the point cloud is sparsified utilising an inverse Gaussian density map $\lambda \in \mathbb{R}^{H \times W}$ following:

$$\lambda = \phi_1 + \phi_2(1 - \frac{ones(\phi_2, \phi_2)}{\phi_2^2} * \frac{(h_2 * (S > \gamma))^2}{max((h_2 * (S > \gamma))^2)}),$$
$$h_2 = \frac{1}{2\pi\sigma_w^2} e^{-\frac{u^2+v^2}{2\sigma_w^2}}, \tag{6.4}$$

where ϕ denotes the weighting parameters. The point cloud is then iterated over, where Eq. 6.5 is repeated until all points are separated by their corresponding λ_a distance threshold:

$$A := \{a \in \chi \mid \exists b \in \chi, b \neq a : dist(a, b) \leq \lambda_a\},$$
$$\chi - A := \{a \in \chi \mid \forall b \in \chi, b \neq a : dist(a, b) > \lambda_a\} \tag{6.5}$$

where, a, b indexes elements in the point cloud set, and $dist()$ gives the Euclidean distance between two points (pixels).

To generate the topological representation of the visible tissue, a Vietoris–Rips complex is then constructed on the sparse point cloud, defined:

$$\Psi := \{d \subseteq \chi \mid \forall a, b \in d : dist(a, b) \leq \epsilon\} \tag{6.6}$$

where ϵ defines the maximum distance threshold. The Vietoris–Rips complex method was chosen over alternatives such as the Delaunay method because of its simplicity and robustness to outliers.

6.2.2 Artefact Detection and Confidence Map Generation

2D simplexes are formed from three points if they are within a predefined distance $\Delta := \{\delta : \delta \in \Psi\}$. For each scan line $\xi_{sl}, sl = [1..width]$, the method will assign 1 for AS and 0 for visible tissue. We define contact-related AS, along a single scan line ξ_{sl}, if the number of overlapping 2D simplexes $\forall\delta \in \Delta$ is less than a threshold τ.

The confidence map is a pixel-wise quantification of the perceptual salience, which represents the resolution and contrast of the detected visible tissue. A high score indicates high-quality tissue contrast and resolution. For every pixel in the image, the number of 2D simplexes that overlap is counted, creating an occurrence map O. The confidence map is constructed by taking the log of the occurrence map, $\hat{M} = log(O + 1)$.

6.3 Resources And Evaluation Setup

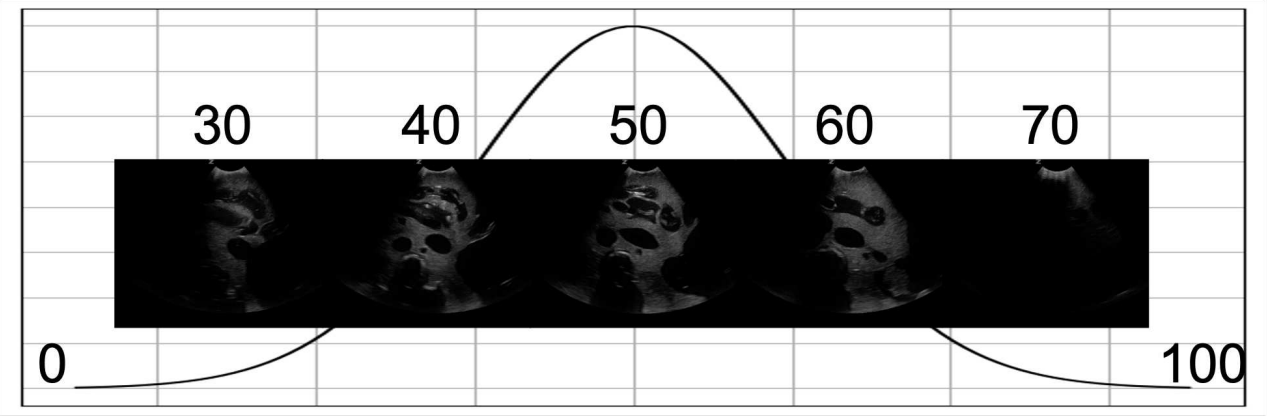


Figure 6.3: The image feedback at different frames from Scan 1 - abdominal. A normal (Gaussian) curve has been used to approximate the confidence behaviour of an ultrasound probe as it tilts, where the optimal imaging is produced when the probe is perpendicular to the tissue surface. This is shown in frame 50, which provides the highest quality visible feedback. As the probe moves away from perpendicularity, image quality reduces in a symmetric, Gaussian pattern. This is evident from the neighbouring frames. Frames 30 and 70, the furthest shown examples from optimal, show the lowest image quality and the least amount of visual feedback. Frames 40 and 60 produce images of slightly reduced quality than frame 50, particularly in terms of sharpness.

6.3.1 Intraoperative Data

6.3.1.1 Data curation

For the AS classification, a dataset containing 51 images from 11 different patients was collected. The images were sourced from routine pseudo-anonymised 2D US scans, acquired in vivo during brain surgery, using a Canon i900 US machine, using both a linear and a convex probe. The data was retrospectively reviewed and annotated by a consultant neuroradiologist experienced in neurooncological IOUS. The images exhibit varying degrees of AS artefacts secondary to poor probe-tissue contact. The study had full local ethical approval from the HRA and Health and Care Research Wales (HCRW) authorities[1]. For standardisation, all convex images were linearised using a standard approach of taking the inner trapezoid and interpolating. All linear

[1]Study title - US-CNS: Multiparametric Advanced Ultrasound Imaging of the Central Nervous System Intraoperatively and Through Gaps in the Bone, IRAS project ID: 275556, Protocol number: 22CX7609, REC reference: 22/WA/0259, Sponsor: Research Governance and Integrity Team (RGIT).

images have the shape 300×600. To reduce the influence of near-field artefacts, the top 100 rows are cropped from all linearised US images. This choice was based on observations from the in vivo dataset and the maximum point where the near-field artefacts, caused by factors that included sheathing, gel, and cranial cavity interaction, ended. This parameter was varied for assessment across all datasets and it was noted that similar performance was achievable in the evaluation, using the tuned parameters, and changing the crop by up to 20% resulted in a $< 1\%$ drop in classification accuracy.

For performance evaluation, three-fold cross-validation was used, the dataset is split into a seen and unseen subset, with the division done on the patient level.

6.3.1.2 Speckle noise robustness evaluation setup

Speckle noise is the most prominent artefact, created by the interaction of signals with biological tissue, which can cause distortion of anatomical details and edges. As proposed in [Jai18], the observed US image I can be defined as the product of the underlying clean image $\hat{I}$ affected by multiplicative ξ_m and additive ξ_a noise, given by:

$$I = \hat{I} \odot \xi_m \oplus \xi_a \tag{6.7}$$

where, it is commonly assumed that $||\xi_a||^2 \ll ||\hat{I} \odot \xi_m||^2$ [ZLG98]. Commonly, the additive term is ignored due to the lesser impact. Assuming sampling from a generalised Gamma distribution (GGD) [MT06] with the probability density function (pdf):

$$p_Z(z) = \frac{\gamma_{ggd} z^{\gamma_{ggd}\nu - 1}}{\alpha^{\gamma_{ggd}\nu}\Gamma(\nu)} \exp\left\{-\frac{z}{\alpha}^{\gamma_{ggd}}\right\}, z \geq 0, \alpha, \nu, \gamma_{ggd} > 0 \tag{6.8}$$

To demonstrate robustness to speckle noise, anatomical edges such as bone or layered tissue is models as a square wave $\hat{I}$:

$$Q(x) = \begin{cases} q_l & \text{if } x < a \\ q_u & \text{if } a \leq x < b \\ q_l & \text{if } x \geq b \end{cases}, 0 \leq q_l \ll q_u \leq 255 \tag{6.9}$$

where a, b are the transition time steps and q_l, q_u are the lower and upper values. Consequently

$$(q_l \odot \xi_m \oplus \xi_a) \ll q_u \quad \text{and} \quad \begin{aligned} |q_l - (q_l \odot \xi_m \oplus \xi_a)| &\ll |q_u - q_l|, \\ |q_u - (q_u \odot \xi_m \oplus \xi_a)| &\ll |q_u - q_l|, \end{aligned}$$

which therefore indicates that artificial structures could only be created by sampling from the extreme tail of p_Z

$$P\left(\bigcap_{i=1}^{n}\{Z_i > \hat{z}\}\right) = \left(\int_{\hat{z}}^{\infty} p_Z(z)\, dz\right)^n, \quad \hat{z} \gg 0, \tag{6.10}$$

which is exponentially unlikely.

Experimental evaluation of the robustness of the algorithm to speckle noise is performed by quantifying the topological feature discrepancies when speckle noise is applied. The speckle noise applied via Eq. 6.7, 6.8, with the ν, γ_{ggd} values varied within a range $\{1, 1.5, 2, 2.5, 3\}$, and sampled from the GGD, using Rejection Sampling [Neu51]. To evaluate the topological difference, the q-Wasserstein distance (with $q = 2$), is used to measure the discrepancies between persistence diagrams, derived from the output point cloud [MBGY14, FCG+21, LCO18]. This measures the topological perturbation under varying noise influence. The q-Wasserstein distance can be defined as $W_q(D, \hat{D}) = \inf_{\varphi: D \to \hat{D}} \left(\sum_{t \in D} ||t - \varphi(t)||_\infty^q\right)^{\frac{1}{q}}$ [MMH11], where φ ranges over bijections between the two persistence diagrams D and $\hat{D}$ (affected by speckle noise). W_q is normalised for interpretability.

6.3.1.3 Classification evaluation and comparison setup

For the evaluation of the AS classification, the proposed method ($Topol$) is compared to a statistical method that uses the sum of the intensity of a scan line ($Baseline_1$), a statistical

method that uses the change in the intensity of the pixels ($Baseline_2$) and a neural network $VGG16$ [SZ15]. The $Baseline_1$ method was motivated by [dSFSY+06] [LKHK18]. The use of the VGG16 architecture was motivated by [MASC22]. $Baseline_2$ is used for the ablation study, the proposed method up to the iterative filtering standard deviation. Confusion matrix analysis is performed, where a positive classification is for AS. The methods are compared in terms of Accuracy (ACC), Specificity (TNR), Sensitivity (TPR) and Precision (PPV). Where: True positive TP, true negative TN, false positive FP, false negative FN, $ACC = (TP+TN)/(TP+TN+FP+FN)$, $TNR = TN/(TN+FP)$, $TPR = TP/(TP+FN)$, $PPV = TP/(TP+FP)$. The methods are ranked using $F_\beta = \frac{(1+\beta^2)\times \text{ppv}\times \text{tpr}}{\beta^2 \times \text{ppv}+\text{tpr}}$, where $\beta = 0.5$ to prioritise precision over sensitivity.

The parameter tuning for Topol is done for each fold using the seen subset exclusively. The best performing set of parameters for $Topol$ - found using a grid search - was near identical across the 3 folds $[f = [2..6], \sigma_u = 1.8, \sigma_w = 20, \phi_1 = 15, \phi_2 = 20, \epsilon = 60]$. With the minor exception of the Gaussian filter where folds 1 and 3 used $\sigma_v = 0.2$, whilst fold 2 used $\sigma_v = 0.3$. The scan line threshold, to determine the AS, was constant for all folds $\tau = 2$. Although the parameters were chosen using a manually defined grid, this process could easily be performed fully automatically using any hyperparameter optimisation method.

$Baseline_1$ individually classifies each scan line through scan line summing, with a fixed threshold $\sum_{i=1}^{300} \xi_{sl,i} < \kappa_{Baseline_1}$. The threshold is determined using a receiver operating characteristic curve (ROC). For folds $[1, 2, 3]$ - $\kappa_{Baseline_1} = [4900, 5000, 5500]$. The ROC process is repeated for $Baseline_2$, where for folds $[1, 2, 3]$ - $\kappa_{Baseline_2} = [325, 350, 350]$. A deep learning method is implemented to gauge generalisability given data limitations. Linear probing was used with a frozen pre-trained VGG16 feature encoder (ImageNet) and trainable classification head. Training is performed for 10 epochs using Sigmoid cross-entropy. $Topol$, $Baseline_1$ and $Baseline_2$, were tuned using ACC.

For the proposed method, the robustness of the parameters is explored by individually perturbing the calibrated parameters from fold 3 and measuring the effect on the corresponding unseen data. For the parameter study, a maximum perturbation of 20% was deemed suitable

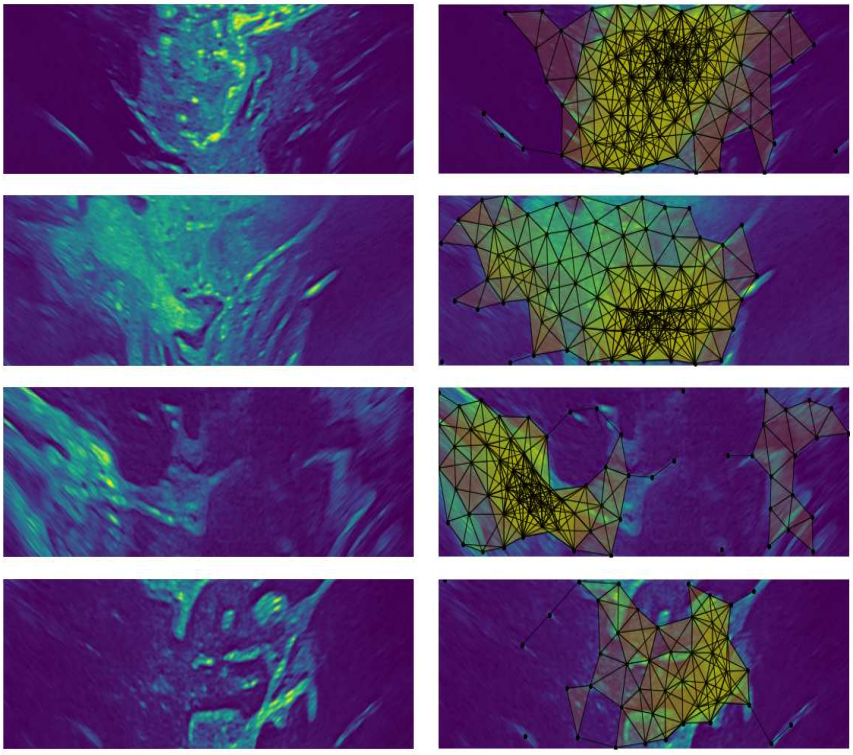


Figure 6.4: Sample simplicial complex outputs. These complexes indicate areas of significant spatial intensity variation within the ultrasound images. Areas where there are dense complexes, correlate with high confidence in tissue visibility; correlating with a confidence map output of higher perceptual quality.

for a valid robustness evaluation. The coefficient of variation $= 100 \times \frac{\sigma_{pert}}{\mu_{pert}}$ is used to measure the impact on performance.

6.3.1.4 Robotic Setup

To extend the AS experiments, a robotic setup is created using the AS output to guide a scanning process - parameters from fold 1 were used. Specifically, the landing of the US probe, then rotation and tilting. The experiments were carried out using a seven DoF DLR Miro [SKB+18], with a GE LOGIQ e US machine and a GE 12L-RS linear US probe. The phantom is created using a mixture of GELITA®GELATINE Type Ballistic 3 with water and glycerine, with olives and blueberries added for echo diversity [MTP07]. The robotic setup is highlighted in Figure 6.5 and Figure 6.6.

Through Cartesian forward kinematics $f_{\text{kin}} : \mathcal{Q} \to SE(3) \mid \boldsymbol{q} \mapsto \boldsymbol{x} = f_{\text{kin}}(\boldsymbol{q})$, the pose of the US

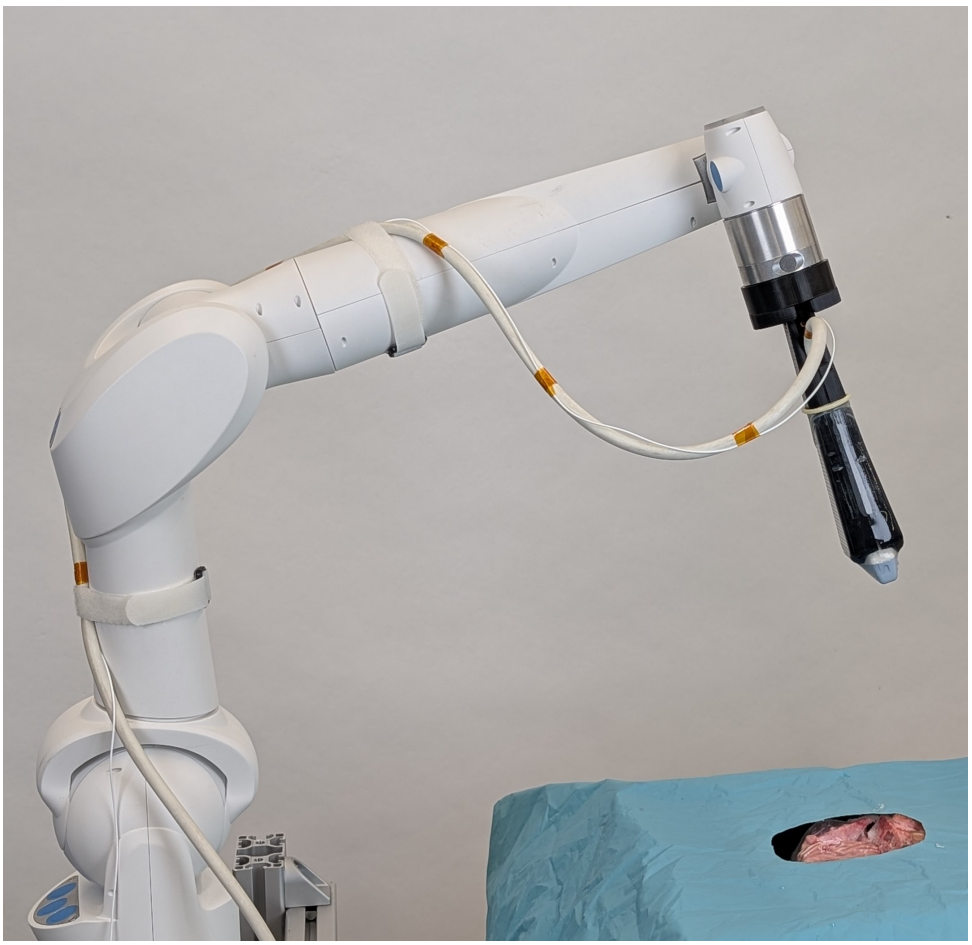


Figure 6.5: Example of MIRO robotic arm setup at the DLR site in Oberpfaffenhofen. In this instance the experiment is being conducted on a porcine brain, for separate work as part of the collaboration.

probe $\boldsymbol{x} \in SE(3)$ can be represented as a function of the generalised configuration coordinates of the robot $\boldsymbol{q}$. A Cartesian impedance controller [Hog85, Kha87] can be constructed to control a robot, by determining desired joint torques $\boldsymbol{\tau} \in \mathbb{R}^n$.

$$\boldsymbol{\tau} = \boldsymbol{J}^{\mathsf{T}}(\boldsymbol{q})[\boldsymbol{K}(\boldsymbol{x}_d - \boldsymbol{x}) + \boldsymbol{D}(\boldsymbol{q})(\dot{\boldsymbol{x}}_d - \dot{\boldsymbol{x}})], \tag{6.11}$$

where $\boldsymbol{x}$ and $\dot{\boldsymbol{x}}$ are the current pose and velocity. This forms a spring-damper system with stiffness and damping parameters $\boldsymbol{K}, \boldsymbol{D}(\boldsymbol{q}) \in \mathbb{R}^{6\times6}$. Enabling control of pose $\boldsymbol{x}_d$ and velocity $\dot{\boldsymbol{x}}_d$ of a US probe. The geometrical Jacobian $\boldsymbol{J}(\boldsymbol{q}) \in \mathbb{R}^{6\times7}$ is given by $\dot{\boldsymbol{x}} = \begin{bmatrix} \boldsymbol{v} & \boldsymbol{\omega} \end{bmatrix}^{\mathsf{T}} = \boldsymbol{J}(\boldsymbol{q})\dot{\boldsymbol{q}}$.



Figure 6.6: Scanning on the home made phantom.

$$m_1 = \frac{1}{width} \sum_{sl=1}^{width} (1 - \xi_{sl}), \quad m_1 \in [0; 1] \tag{6.12}$$

$$m_2 = \frac{1}{width/2} \left(\sum_{sl=width/2}^{width} \xi_{sl} - \sum_{sl=1}^{width/2} \xi_{sl} \right), \quad m_2 \in [-1; 1], \tag{6.13}$$

Two metrics are defined to measure the coupling - m_1 is the coverage of contact, m_2 is the symmetry of contact. These metrics can guide the scan by defining desired values $m_{1,d}$ and $m_{2,d}$ and mapping the error to a desired Cartesian velocity.

$$\dot{\boldsymbol{x}}_d = \begin{bmatrix} \boldsymbol{v}_{d,m} \\ \boldsymbol{\omega}_{d,m} + \boldsymbol{\omega}_{d,t} + \boldsymbol{\omega}_{d,r} \end{bmatrix} = \begin{bmatrix} (m_{1,d} - m_1) \begin{bmatrix} 0 & 0 & k_1 \end{bmatrix}^\mathsf{T} \\ (m_{2,d} - m_2) \begin{bmatrix} k_2 & 0 & 0 \end{bmatrix}^\mathsf{T} + \boldsymbol{\omega}_{d,t} + \boldsymbol{\omega}_{d,r} \end{bmatrix}, \tag{6.14}$$

assuming the US probe scan axis is aligned with the y-axis of the tool frame, with two gains

$k_1 = 0.005$ $^{\text{m}}/_{\text{s}}$ and $k_2 = 0.05$ $^{\text{rad}}/_{\text{s}}$. The translation velocity is defined by $\boldsymbol{v}_{d,m}$, while the rotation velocities are defined by $[\boldsymbol{\omega}_{d,m}, \boldsymbol{\omega}_{d,t}, \boldsymbol{\omega}_{d,r}]$. The integration of Eq. (6.14) gives the desired pose $\boldsymbol{x}_d$. If no trajectory is defined ($\boldsymbol{\omega}_{d,t} = \boldsymbol{\omega}_{d,r} = [0, 0, 0]$), so, for example, when the probe is landing or maintaining perpendicular contact. Using $\dot{\boldsymbol{x}}_d$, the robot will automatically adjust the position of the probe along the longitudinal axis until $m_1 = m_{1,d}$, and adapt the orientation until ($m_2 = m_{2,d} = 0$. To perform rotation and tilt, $\boldsymbol{\omega}_{d,t} = \begin{bmatrix} 0 & * & 0 \end{bmatrix}^{\mathsf{T}}$ and $\boldsymbol{\omega}_{d,r} = \begin{bmatrix} 0 & 0 & * \end{bmatrix}^{\mathsf{T}}$ accordingly, where $*$ is defined by the operator.

6.3.2 Medical Phantom Data

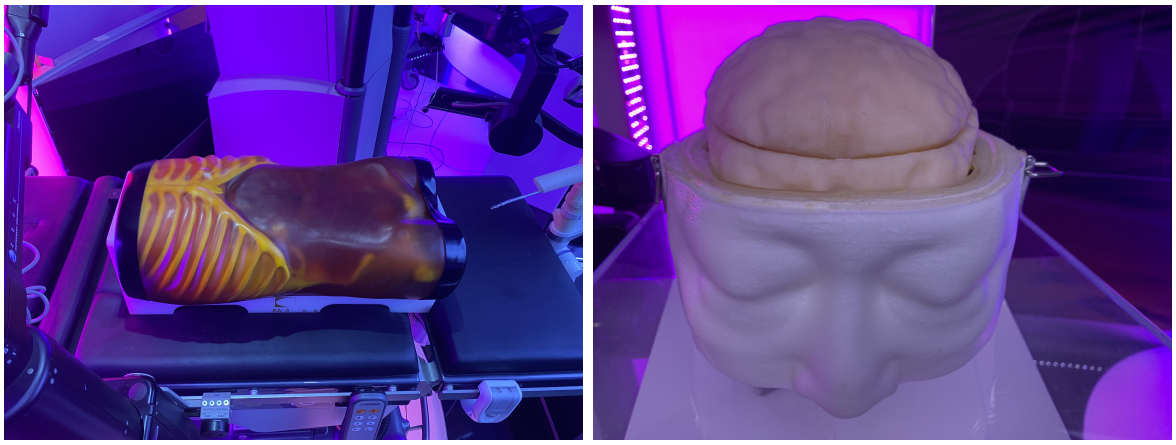


Figure 6.7: The FAST / acute abdomen and Kezlex brain phantom.

6.3.2.1 Data curation

Controlled evaluation of confidence is difficult, as optimal scans may produce subpar images due to unavoidable factors such as forced positioning - for example, when scanning within a resection cavity or when there is obstruction of the area of interest due to the presence of bone or critical tissue. Data was collected for relative evaluation, this involved tilting the probe from -90° > 0°(perpendicular) > +90° with respect to the normal vector of the tissue surface. Data was captured using a Zonare Z.one US system with a C4-1 transducer. Four scans were taken using the FAST / acute abdomen Phantom. Scans 1 and 2 from the abdomen, scan 3 from the rib cage, and scan 4 from the kidney. The two phantoms are shown in Figure 6.7. A fifth scan

was taken using a Kezlex brain phantom. Frames were extracted at $10Hz$. For generalisation evaluation of AS classification, 449 frames are extracted and annotated.

6.3.2.2 Confidence map evaluation and comparison setup

The proposed method is compared with the *Randomwalk* [KWKN12a] and *DAG* [HCG20] methods, using their official codes and default parameters. The *Topol* parameters tuned on fold 1 of the AS classification were used for this experiment. Due to the way the data was collected, the target confidence vector is defined as a normal distribution across the frames. Based on [VZE+16], $mean(\hat{M})$ is used to determine the confidence in the perceptual salience of the visible anatomy per frame. The evaluation is performed using the normalised root mean squared error, $(NRMSE) = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(T_i-\hat{T}_i)^2}}{\max(T)-\min(T)}, \hat{T}_i = mean(\hat{M}_i)$, where T is the target confidence vector, the black dotted lines in Figure 6.10, fitted to the max/min of $\hat{T}$; n is the frame. Figure 6.3 is from Scan 1. The best image is the middle frame, and the black curve is representative of the assumed confidence trajectory. To confirm this, annotations were made to highlight the optimal frame range, highlighted with the red dotted lines in Figure 6.10.

6.3.2.3 Classification generalisability

To test the generalisability and performance of each algorithm on out-of-distribution data, AS scanline classification experiments were extended to phantom data (different machine, anatomy and image texture). Each method was evaluated three times, using the three sets of parameters calibrated for the three folds described in Sec. 6.4.2, and the average results across the three folds were recorded.

6.4 Results and Discussion

Table 6.1: Scan line classification results. A positive sample corresponds to a scan line affected by AS. Highlighted in grey is the best score for each metric. The results have been averaged over the 3 folds.

3F Avg	Algo	ACC	TNR	TPR	PPV
Seen/Train	$Baseline_1$	0.87	0.88	0.86	0.86
	$Baseline_2$	0.88	0.91	0.85	0.90
	$VGG16$	0.90	0.92	0.90	0.90
	$Topol$	0.89	0.95	0.83	0.93
Unseen/Test	$Baseline_1$	0.84	0.81	0.89	0.79
	$Baseline_2$	0.87	0.85	0.87	0.83
	$VGG16$	0.83	0.82	0.84	0.80
	$Topol$	0.87	0.90	0.84	0.87

Table 6.2: The rankings (R#) of the methods using the F_β scores. The results are a product of averaging across the 3 folds.

3F Avg	R1	R2	R3	R4
Seen/Train	$Topol$	$VGG16$	$Baseline_2$	$Baseline_1$
	0.908	0.900	0.890	0.860
Unseen/Test	$Topol$	$Baseline_2$	$VGG16$	$Baseline_1$
	0.864	0.838	0.808	0.808

Table 6.3: The rankings (R#) of the methods' performances on the generalisability test on the probe pose analysis data, using the F_β scores. For all scans, $Topol$ achieves the best rank.

	R1	R2	R3	R4
Scan 1	$Topol$	$Baseline_2$	$Baseline_1$	$VGG16$
	0.923	0.842	0.807	0.694
Scan 2	$Topol$	$Baseline_2$	$Baseline_1$	$VGG16$
	0.920	0.883	0.789	0.740
Scan 3	$Topol$	$Baseline_2$	$Baseline_1$	$VGG16$
	0.833	0.789	0.745	0.377
Scan 4	$Topol$	$VGG16$	$Baseline_2$	$Baseline_1$
	0.891	0.847	0.797	0.753
Scan 5	$Topol$	$VGG16$	$Baseline_2$	$Baseline_1$
	0.830	0.777	0.686	0.643

The evaluation was carried out primarily using a Python implementation; results are shown in Figure 6.4. The method was also implemented with minimal optimisation using C++, which ran at 20Hz, indicating real-time executability.

Table 6.4: The method's NRMSE results are shown for each scan. In all cases, *Topol* produces the best score.

Method	Scan 1	Scan 2	Scan 3	Scan 4	Scan 5
Topol	0.18	0.13	0.28	0.14	0.11
Randomwalk	1.85	0.94	4.79	0.73	0.87
DAG	2.40	4.96	6.08	5.12	4.15

6.4.1 Noise Robustness

The q-Wasserstein distances, after applying the set speckle noise $\{1, 1.5, 2, 2.5, 3\}\times\{1, 1.5, 2, 2.5, 3\}$ for $\nu \times \gamma_{ggd}$, measured an average result of 0.097. This result gives evidence of the robustness of the topological construction under varying speckle noise. The topological differences are largely uniform across all distributions.

6.4.2 Acoustic Shadow Classification

The results are presented in Tab. 6.1, 6.2. Outputs are shown in Figure 6.8. Firstly, it can be determined that *Topol* is the most robust method for distribution shift with more consistent performance between seen and unseen data. The results produced by $Baseline_1$ show that overreliance on the raw pixel intensities leads to oversensitivity. By producing results superior to $Baseline_1$, $Baseline_2$ justifies the iterative filtering process, demonstrating that identifying areas of dynamic change is a strong indicator of where there is AS. $VGG16$ shows achieves inadequate generalisation. From the F_β ranks, it is conclusive that *Topol* produces the ideal behaviour.

6.4.2.1 Generalisability Study

The results of this experiment are shown in Tab. 6.3. For all scans, *Topol* achieves the best performance. The ranking of $Baseline_2$ again highlights the advantage of using the intensity change rather than the raw intensity. The inconsistent performance of $VGG16$ highlights the unreliability of deep learning methods when only small datasets are available.

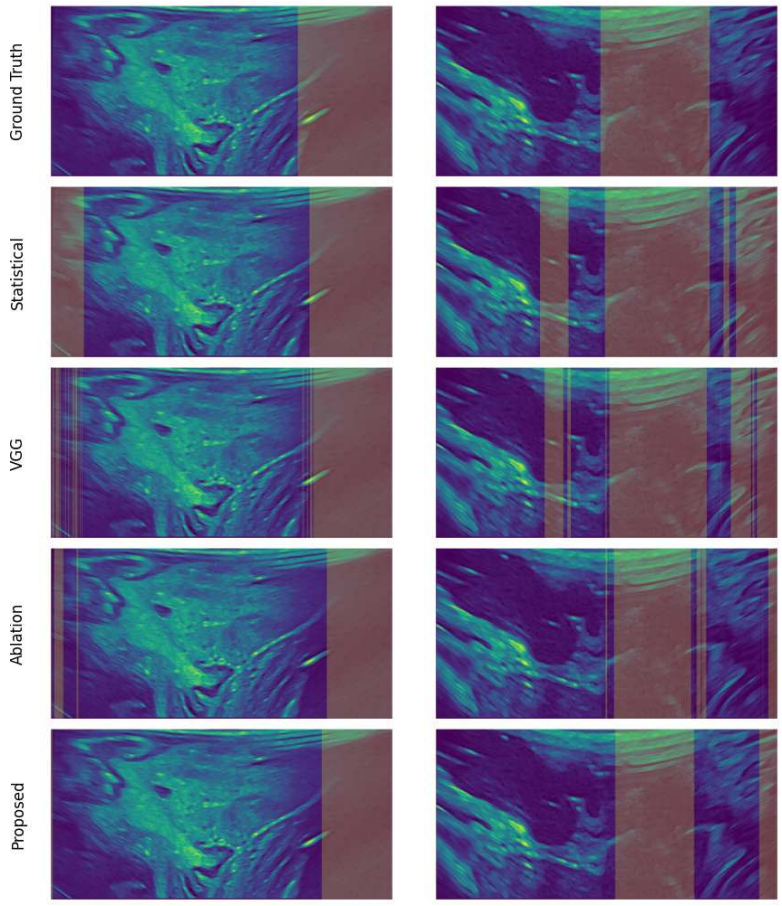


Figure 6.8: AS detection on two US images. The top row shows the annotation, the second is from $Baseline_1$, the third is from $VGG16$ and the bottom is from Topol. The orange areas are detected AS.

6.4.2.2 Parameter Perturbation Study

The results of this experiment are shown in Figure 6.9. It is observed that the overall performance is only marginally affected by individually drifting the parameters, with most metrics remaining above 0.8. The coefficient of variation values (%) of the performance drift are shown in Tab.6.5. This is evidence of the robustness of the design of the algorithm, as the performances do not vary largely under parameter uncertainty. We note that the γ value that was decided was initially based on intuition from the iterative filtering process. In particular, calculating the overall standard deviation $\sqrt{5} \times 1.8 \approx 4.02$ yields a value close to γ=4. This result suggests a potential correlation between the iteration and the optimal threshold choice.

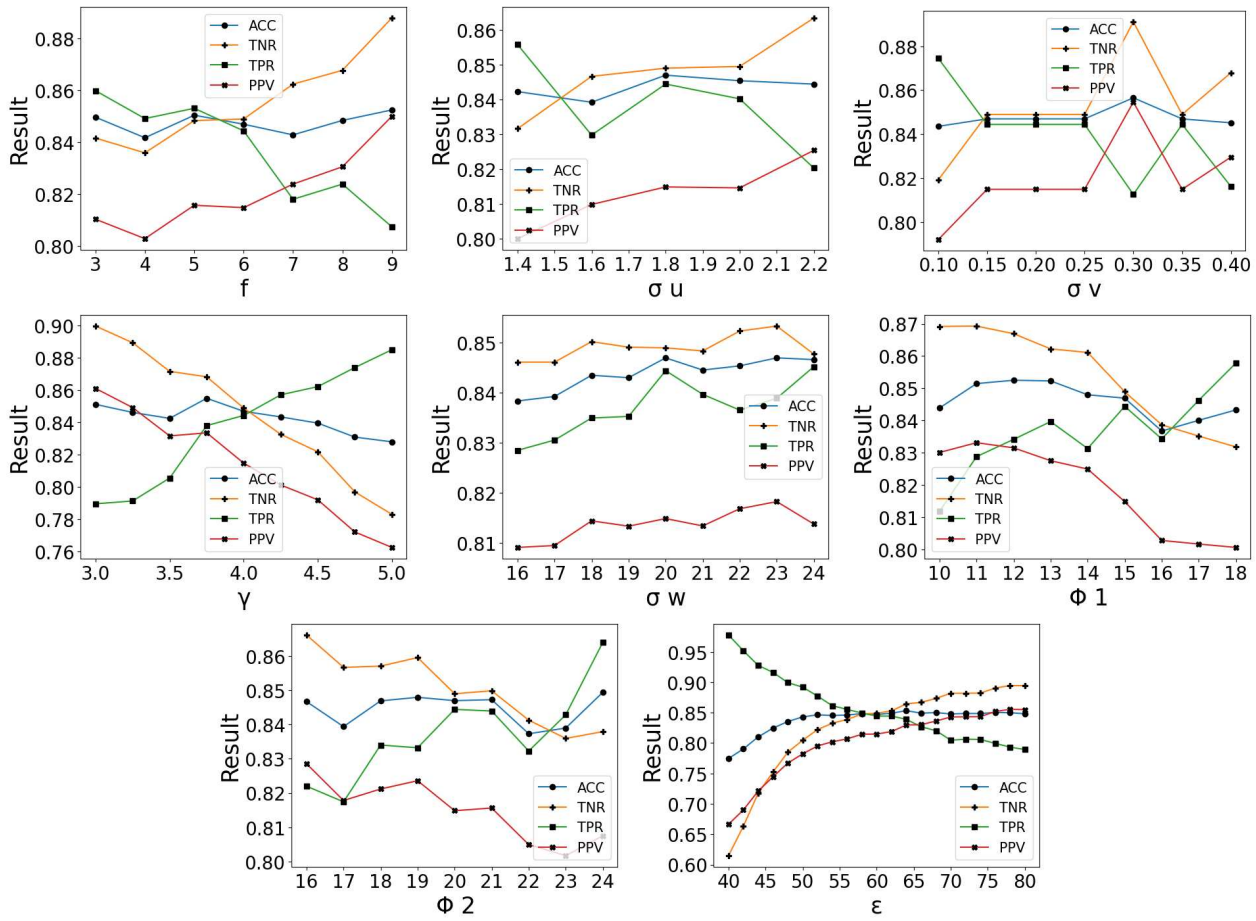


Figure 6.9: Graphs displaying the effect of individually varying the parameters on the performance of the whole system - using the unseen fold 3 data. blue lines = ACC, orange lines = TNR, green lines = TPR and red lines = PPV.

6.4.3 Robotics results

The sequence of robotic action is first the landing and stabilisation of the probe on the phantom surface, followed by a ±90° rotation, followed by a ±30° tilt. The real homogeneous transformations of the robot base, as well as the orientation commands of the probe, are shown in Figure 6.11. Of the translation components, Z is of greatest interest, as the scan was held on a single spot, and Z was the main translation for correction. The real homogeneous transformations matrices show that the robot correctly executed the motion without large deviation. The AS results for $[m_1, m_2]$ are shown in Figure 6.12. It can be determined that AS analysis enabled the control framework to successfully anchor the contact to the desired coverage and symmetry.

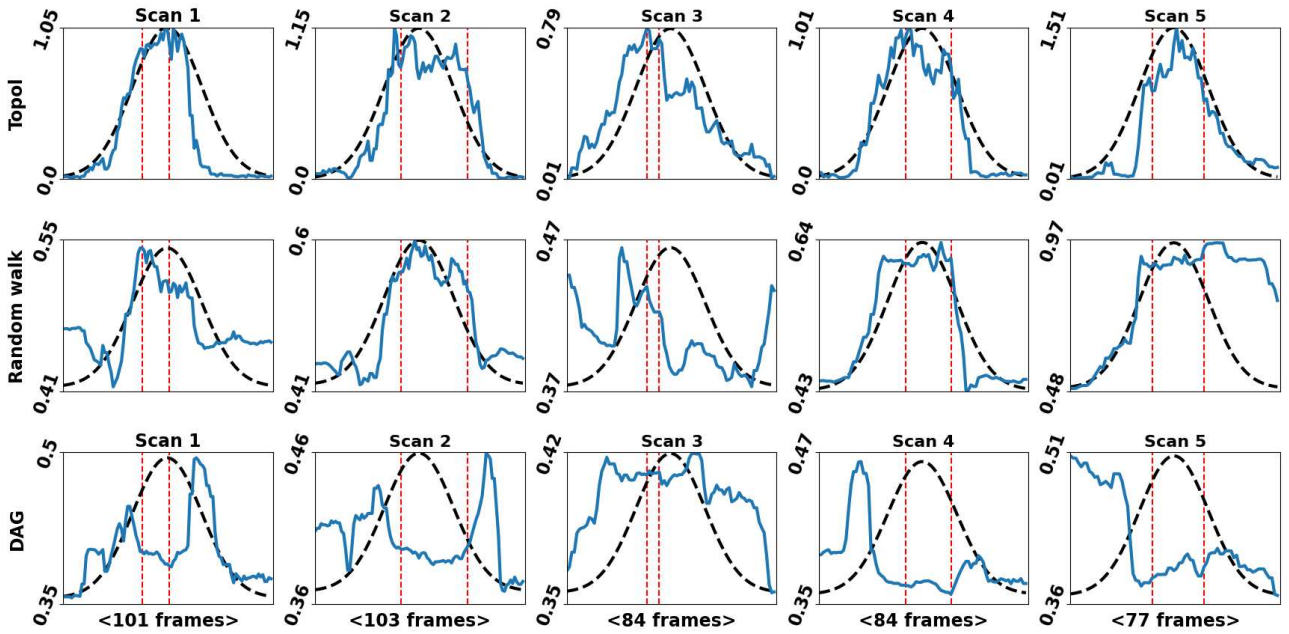


Figure 6.10: The per frame average confidence vector outputs - x axis = frames and y axis = average confidence. The blue lines show the method's outputs, *Topol* top, *Randomwalk* middle and *DAG* bottom row. The black dotted lines are the scaled normal curve. The dotted red lines are where the annotators perceived the best images. What can be observed is that the output of *Topol* follows the desired trajectory, closer than the other methods.

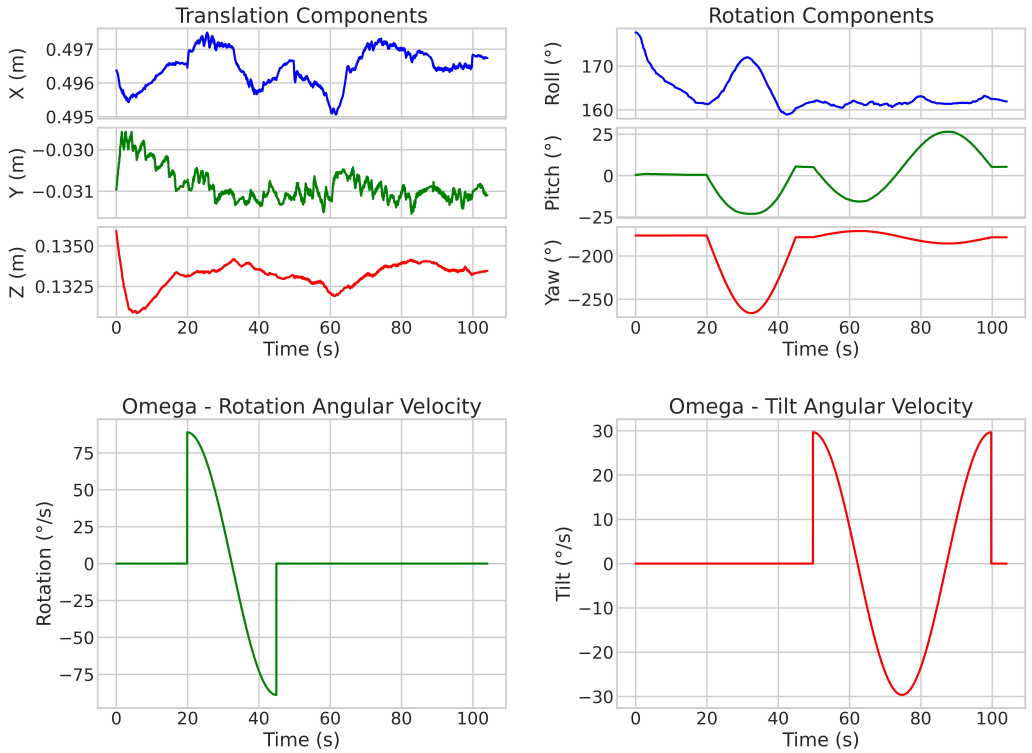


Figure 6.11: Robot base homogeneous transformation matrices and desired rotation and tilt commands.

Table 6.5: Parameter perturbation, performance drift measured via coefficient of variation values (%).

Parameter	ACC	TNR	TPR	PPV
f	0.431	1.928	2.204	1.749
σ_u	0.320	1.191	1.455	1.011
σ_v	0.455	2.380	2.290	2.151
γ	0.982	4.499	3.986	3.925
ϕ_1	0.622	1.683	1.453	1.579
ϕ_2	0.512	1.150	1.550	1.031
σ_w	0.358	0.277	0.635	0.347
ϵ	2.270	8.663	5.986	6.306

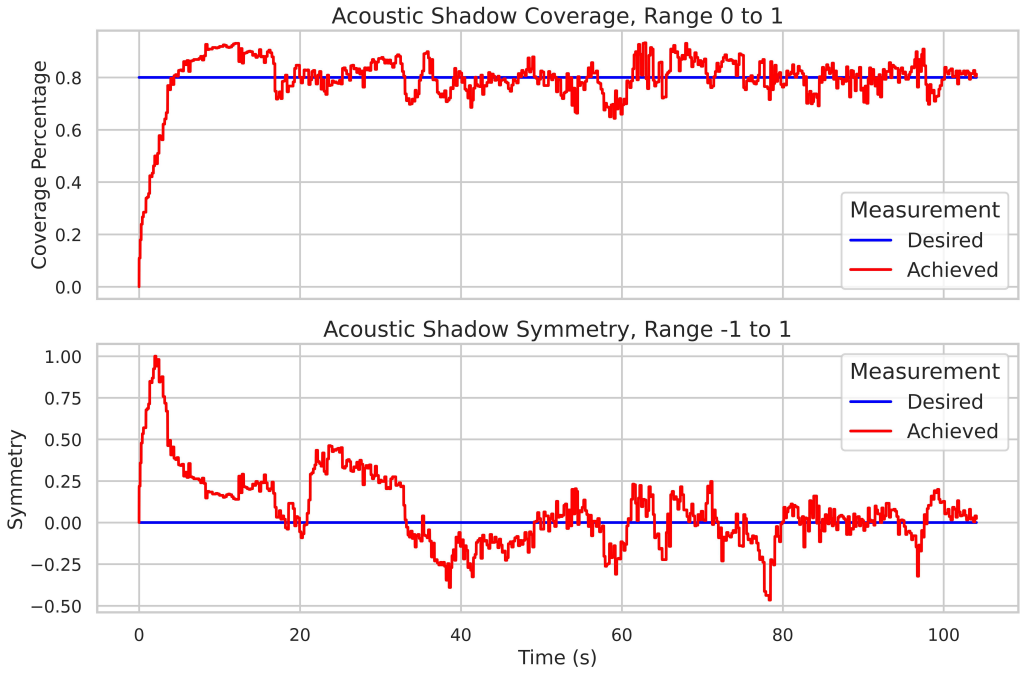


Figure 6.12: Acoustic shadow behaviour and optimisation for robot scan.

6.4.4 Confidence Map Behaviour

Figure 6.10 shows that for all scans, the behaviour of *Topol* is closer to what is expected. The range of $\hat{T}$ values produced by *Topol* can also be considered more intuitive than *Randomwalk* and *DAG* as the confidence starts near 0 and peaks around 1. Quantitative analysis is also performed and shown in Tab. 6.4, produced by measuring the NRMSE of $\hat{T}$ (blue line) versus T (black dotted line). For each scan, the proposed method produces a significantly better NRMSE score than the other methods.

6.5 Conclusion

This chapter introduces a novel method for evaluating the quality of US feedback. An evaluation of AS scan line classification and confidence behaviour has been performed. In both cases, the proposed method outperforms the compared methods. Further experiments show that this method is capable of guiding a robotic arm in performing an autonomous scan. The explicit evaluation of US probe-tissue coupling is vital for ensuring safe scanning. This method provides a novel perspective on how this can be approached, creating an avenue of future work which can be beneficial for both robotic US and supporting trainee operators.

A limitation of our evaluation method - and a potential direction for future work - lies in how we model speckle noise. Speckle artefacts are generated by drawing independent, pixel-wise samples from a GGD and applying them multiplicatively (optionally with an additive term) to the B-mode images. This approach effectively replicates the first-order envelope properties of speckle noise - dominant factor for our filter's performance - of US speckle by modulating the amplitude distribution of individual pixels and is widely adopted in US image processing [MT06, VSFAPMF12]. However, it does not reproduce second-order statistics, i.e. the spatial correlations that affect grain size and texture. Exploring these second-order models, and potentially integrating them directly into methods like ours, therefore remains an interesting avenue for future work, especially for challenging anatomy such as brain tissue.

Part III

Resource Leveraging

Preamble

This Part of the thesis contains work that aims to directly tackle problems within one of the central themes that run throughout this thesis, data limitation. Through work within the JADs project, two areas of critical resource limitation resulted in impediment of the development of technical solutions for automated IOUS. The first is the subject of 3D vision, and the second is the subject of US realistic phantoms.

Although robotic platforms for US can be developed without any requirement for 3D information of the surgical scene, as seen in Part 2. If the issue of additional hardware and setup arrangement can be minimised, there would be limited justification to exclude the additional source of information, specifically for prescan modelling; during the scan the tissue surface will be mostly occluded by the probe. 3D vision is an extremely well-trodden research area that has many demonstrations of successful translation to real-world application. There are many avenues for solving this problem, for example, using active sensors such as LIDAR, utilising multiple cameras, graphics methods such as Gaussian splatting or neural radiance fields, etc. For the first Chapter in this Part, I will explore 3D vision for single, stereo cameras. This was chosen as these cameras are the most common sensor type in clinical practice along with monocular cameras. In addition, they have much higher data capture resolution than alternatives such as LIDAR. Most importantly, multi-view camera-based deep learning methods have demonstrated far superior performance and viability for 3D tasks compared to alternative sensors and algorithmic methods. The caveat, again, is the requirement for data; where even generic, clinical tissue 3D public data is extremely limited. In my first work in this Part, I explore a strategy for collecting clinical task-related datasets with ground-truth 3D information, and propose a method for neural network training given a dataset that, by regular deep learning standards, would most likely be limited/suboptimal. I hope that this work encourages the collection of task-specific data for the development of 3D vision methods with more niche applications, such as brain surgery. In relation to Part 2, this methodology could offer a solution to develop a 3D vision component of a robotic system to assist in the automation of the landing of an US probe onto brain tissue.

Thorough validation of clinical systems is essential to demonstrate performance and ethical compliance, particularly for automated analysis and robotic frameworks. Acquiring ethical permission to conduct experiments in vivo and using ex vivo tissue that contains the relevant pathological tissue is a significant logistical hurdle. Often, early experiments are conducted using participants (if non-invasive) or using synthetic medical phantoms. For our robotic work, the latter was the only feasible option to validate our system. I believe that there are two main factors that make synthetic phantoms usable for experimental development and evaluation, cost and quality/realism. As experiments will be numerous and the budget of different research groups is highly variable, phantoms need to be affordable. Paramount to the usefulness of a phantom is it's quality and realism of US image feedback. Unfortunately, commercially available synthetic phantoms tend to be expensive and provide noticeably simpler and textureless US imagery, compared to living real tissue. This is particularly the case when considering the improved imagery provided in state-of-the-art US machines, and for brain tissue, which is highly complex. In my second work in this Part, I provide an inexpensive method for creating realistic brain phantoms, containing pathological tissue. This, I hope, provides a bridge between evaluation using computational simulations and unrealistic phantoms with real in vivo experiments. A useful resource for supporting the clinical translation of technical systems. In relation to Part 1, this phantom could be used to help supplement datasets for deep learning training. In relation to Part 2, the evaluation of robotic systems for pathological tissue discovery and mapping in brain IOUS could be conducted using this phantom.

Chapter 7

Learning Structured Light Projections To Improve Stereo Disparity Learning From Limited Data

7.1 Introduction

Recently, it has been shown that when large datasets are available, deep learning approaches define the state-of-the-art in 3D scene reconstruction [ZSZ+20, LJBB20]. This is fundamentally due to a neural network's ability to learn more complex representations of image data than can be handcrafted. However, the coupling of data volume and performance is an issue, particularly for domains that have limited data availability such as surgery [HRRM18, WKH+20]. Capturing large amounts of depth information for surgery, especially Minimally Invasive Surgery (MIS), is laborious, due to issues with hardware constraints and the difficulty of dealing with the tissue; primarily issues with the deformation of the tissue which can disturb information capture. This problems is reflected in the SCARED dataset [AMW+21], which is the largest annotated depth dataset for surgical scenes, but only contains 45 unique and complete depth images. Training on small datasets complicates the development of networks [QL20, BI21], due to the risk of overfitting [Yin19]. To obviate ground truth, stereo self-supervision [GAB17] ap-

proaches have been developed; learning disparity by training to warp the left image to become the right and vice versa. However, results are still considerably worse than what is achieved by supervised networks [USS+17]. Accurate 3D reconstruction can provide surgeons a tool for surgical planning and guidance [HMWL+21]. Therefore, strategies for overcoming data limitation are desperately needed. Otherwise, the development of accurate neural networks for surgery is unachievable.

Structured light is currently the most dense and accurate approach for creating ground truth information for depth datasets. Example datasets that were created using structured light include NYU [SF11], Middlebury Stereo [SS03b, SHK+14] and SCARED. Structured light is the projection of patterns into a scene, which when captured by an imaging camera, allows for depth recovery through analysis of the pattern distortions [SPB04]. Commonly, these pattern projection images are used exclusively for ground truth depth generation. Once depth has been captured, the pattern projection images are discarded afterwards. However, error will occur in the conversion process due to difficulties relating to the environment, surface properties and the hardware [RMS19, JWA17, GAVN11, SS03a]. Primarily, errors will occur at the pixel classification stage which for example can be caused by reflections or hardware malfunction. This means that the information within the structured light images and the generated depth maps are not the same.

In this chapter, we are proposing a unique approach to depth estimation. Firstly, we show that it is possible in itself to teach a neural network to be able to artificially project structured light patterns into a scene. Which is indirectly learning to perform disparity estimation. To the best of our knowledge, this is the first use of a neural network for the learning of the projection of structured light patterns into a scene. Secondly, we show that the dual training of direct disparity regression and structured light projection, via multi task learning, enhances network training and improves disparity learning without increasing the number of parameters nor requiring the collection of additional ground truth data.

7.2 Related Work

7.2.1 Active Sensors

Structured light can be used to produce unique pixel codes for each stereo image, which enables the use of simple matching techniques for the stereo disparity calculation. This approach can provide dense and accurate depth maps for a given stereo image. However, structured light requires a controlled environment, multiple hardware and sometimes temporal variation. Certain structured light patterns require sequential projection of different patterns over time, such as binary/gray code which is what this chapter works with. The control of the environment, when performing structured light projection, is important for good pixel classification. Robust classifiers are normally required to prevent incorrect classification [SPB04]. Theses limitations have large ramification for setups designed for real time performance in dynamic environments. Our proposed method for structured light reconstruction removes the need for the additional hardware and the temporal requirement.

7.2.2 Depth estimation

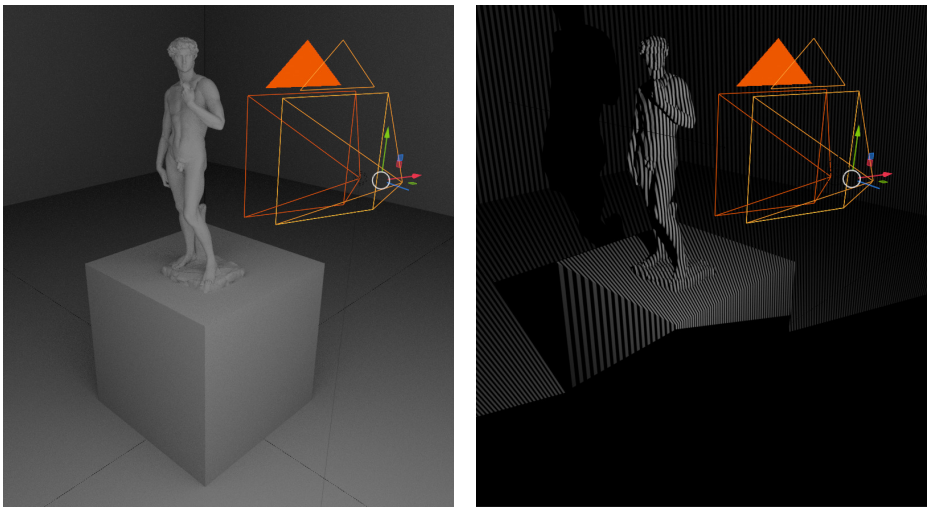
The state-of-the-art for 3D reconstruction is defined by deep learning approaches that directly regress depth values. The general consensus within the deep learning community is that a neural network will discover a better way to solve the stereo matching problem when given free roam in an end-to-end setup where there is limited human supervision. Examples of these networks could be, RAFTStereo [LTD21] or PSMnet [CC18a] for stereo and DORN [FGW+18] for monocular. The prevalent issue with these networks is the requirement of large training datasets. In the papers mentioned, the networks are pre-trained on scene-flow which contains approximately 60k images, then fine tuned on popular sets like KITTI. Although generalisability has been shown to be largely acheivable, the further the data similarity is to what has been trained on, the worse the performance. This poses a large hurdle when the 3D reconstruction task is for specific and niche tasks. In this chapter we show that recycling structured light

images, which is data commonly acquired to produce depth ground truth, can improve the performance of disparity estimation techniques especially when data is limited.

7.2.3 Multi Task Learning

Multi task learning (MTL) has already been proposed for depth/disparity estimation purposes. In previous works, the training of depth/disparity, has, for example, been combined with the training of semantic segmentation or instance segmentation [SK18, KGC18]. These works have shown that it is possible to achieve an improved depth/disparity performance in comparison to when trained as a single task. However, as has been identified in previous literature [SZC+20], MTL does not guarantee improved performance and balancing the combination of tasks and the combination of weights of the loss terms is difficult. In this work we show that it is possible to use structured light reconstruction in a MTL framework to improve disparity learning. Where the ground truth requires no annotation.

7.3 Method



(a) The scene lit with camera light source only.

(b) The scene lit with projector only.

Figure 7.1: Demonstration of the virtual environment. The virtual camera is represented in orange and the projector object is represented in yellow.

In this section we introduce our stereo disparity estimation techniques. Firstly, we propose the first deep learning model to learn the projection of structured light patterns onto a scene. The goal of this model is to show that it is possible to accurately learn structured light patterns that respect the stereoscopic perspectives and verify that structured light information can be used for 3D reconstruction purposes and should not be discarded after ground truth information has been generated; which is what is most commonly done. Secondly, we propose a novel MTL method for disparity estimation, showing that MTL can enhance disparity learning by dual training on structured light.

7.3.1 Dataset generation

Virtual dataset: Due to the lack of any publicly available datasets containing structured light patterns, we had to create our own dataset. A simulated environment was developed using Blender [Com18] to enable automatic generation of virtual scenes containing cultural objects, as shown in Figure 7.1. The VisionBlender add-on [CTL+20] was used to generate the ground truth information.

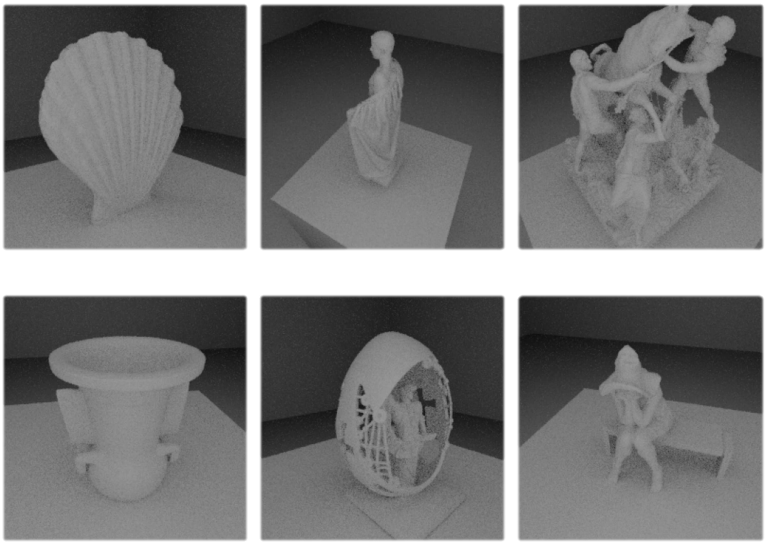


Figure 7.2: Sample objects used to create the virtual dataset.

The virtual environment is simple, consisting of a cubic room of volume 1 $[m^3]$ and a central podium of size 10 $[cm^3]$, which is used to hold different objects. All the surfaces in the whole virtual scene are grey and there is no texture, including the surfaces of the objects. The projector [Sch21] was rigidly attached to the camera object with a rotation of $R = 1.5°$ and a translation of $T = [0.02m, 0.00m, 0.02m]$. 800 cultural artefact objects, were used as the diversity factor in this dataset. The objects contain complex 3D surfaces and a smooth texture, making them challenging to reconstruct using existing 3D reconstruction approaches. All the objects were downloaded from the "Scan The World" project which is hosted by MyMiniFactory [Bec19] - sample objects show in Figure 7.2. The dimensions were resized using a uniform distribution $U(0.03m, 0.1m)$. For each object, 10 stereo image-pairs were captured from different view points. The camera is always looking towards the geometrical centre of the object. The distance between the camera and that geometrical centre is sampled from $U(0.03m, 0.1m)$. Respecting these two constraints, the pose of the 10 viewpoints are chosen at random.

For each viewpoint, the stereo images and the associated ground truth data were generated. The ground truth includes the depth/disparity maps and stereo images without and with 8 projected binary patterns. For each projection, the number of vertical lines equals $2^n, \forall n \in [1..8]$, where n is the projection number - Figure 7.3 displays a full set of structured light patterns. The resolution of the captured images are 256×256. This dataset contains approximately 8000 images.

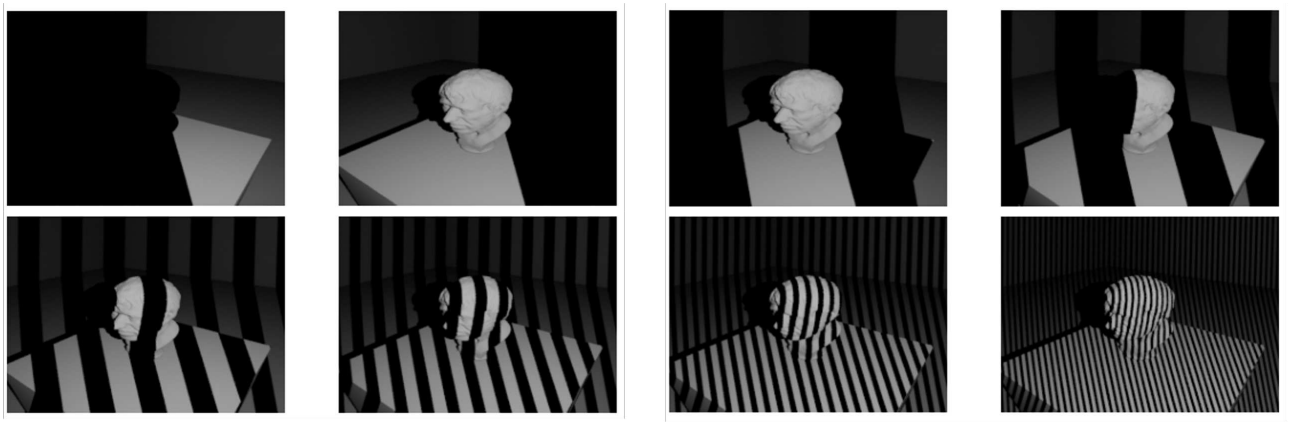


Figure 7.3: Visualisation of the binary code structured light patterns, before binarizing. The images are from the left camera.



(a) The capture of a sheep's liver using the dVRK laparoscope.

(b) An example image captured which contains multiple ex vivo tissue.

Figure 7.4: Shown in (a) and (b) is the setup for the medical dataset collection. Image (a) shows an example of tissue captured using the dVRK laparascope. (b) does not contain projection patterns and would be fed as an input to the neural network.

Medical dataset: A medical dataset was created using the da Vinci Research Kit (dVRK) [KCD+14] to extend the evaluation of the model onto real images. The dVRK was chosen as it is the research model of the surgical system that is used in clinical practice and commonly used for research in this field.

During collection, a 3d printed turning table was used to hold objects that included medical phantoms and ex-vivo organs from sheep, cows and chicken. This allowed for the the capture of 8 unique poses/viewpoints of the contents of on table, where we the set of all rotations as a keyframe. The camera was placed with a viewpoint close enough to the tray such that the

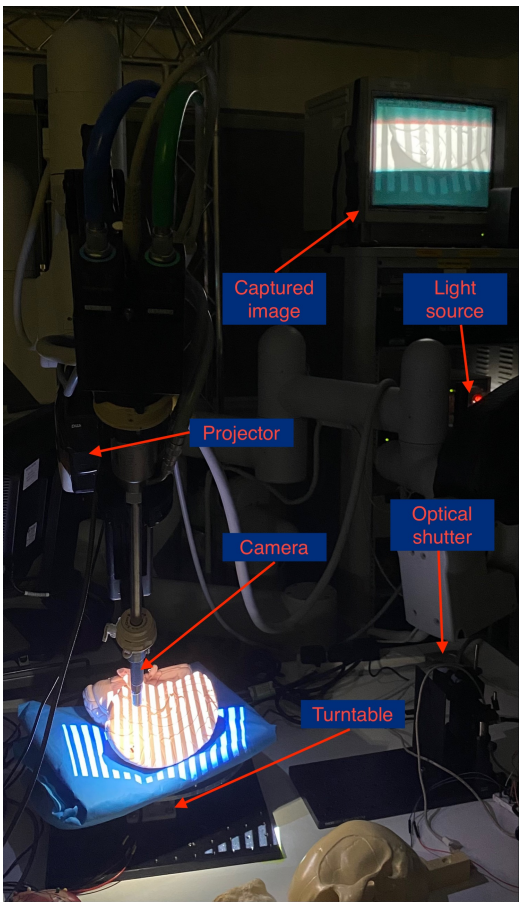


Figure 7.5: The medical dataset image capture setup.

overlap between each pose is minimised. A Raspberry Pi was used as the central controller for the automation process. Sending commands with SSH, the Raspberry Pi captures images via the da Vinci central computer. Structured light patterns were projected using an Optoma projector, which was controlled with a Python script. The projector was rigidly attached to the laprascope by hard fixing it onto the static robotic arm that also held the camera. A Dynamixel servo motor was used to rotate the turntable. The motor was controlled using the ROBOTIS Dynamixel SDK. The laproscope was calibrated using MATLAB's Stereo Camera Calibrator App, where 10 poses of a 21x20x5mm checkerboard was used for the calibration, as shown in Figure 7.7. Gray code was used for this dataset with 9 patterns. The decoding was performed using the three-phase algorithm [XHE22]. For this method, the inverse of each Gray code pattern was additionally projected to refine the line boundaries. A three-phase modulation, shifting of the eighth pattern twice, was captured to generate an uncertainty map within the image. Using the binarised patterns, the depth and disparity is then calculated using decimal



Figure 7.6: Examples of ex vivo used.

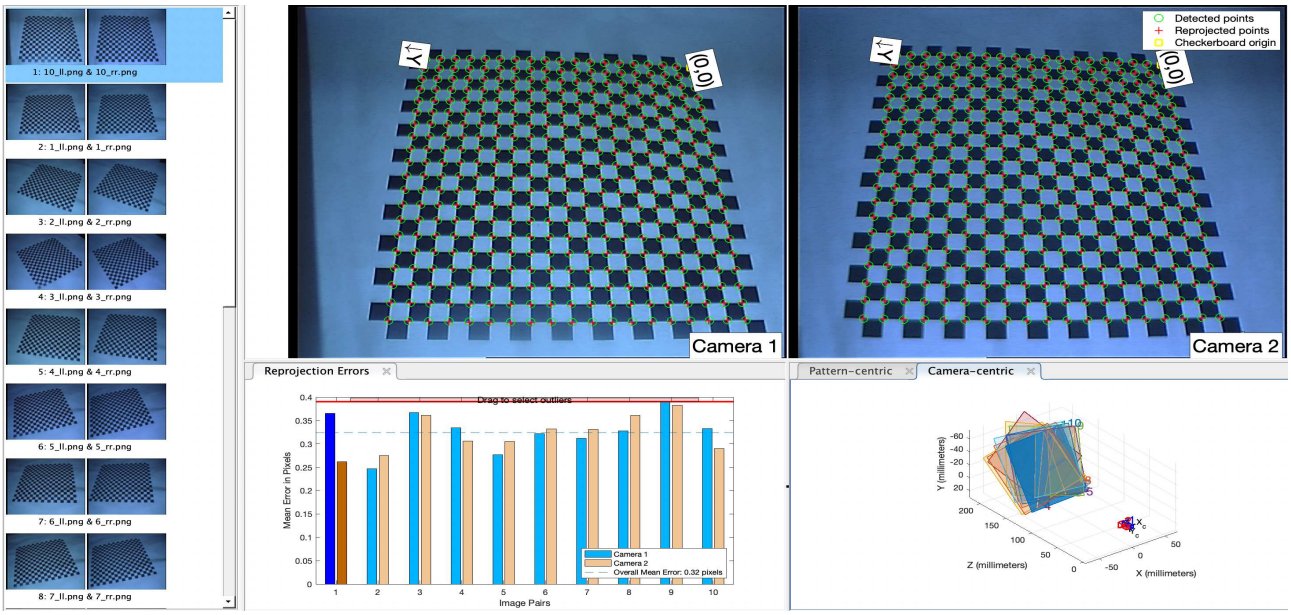


Figure 7.7: The calibration of the laprascope using MATLAB's Stereo Camera Calibrator app.

matching along the epipolar lines. The collected dataset contains 57 keyframes and 379 unique images. Figure 7.5 and 7.4 shows the setup and an example image which would be fed as the input to the neural network (an image without projection patterns).

7.3.2 Structured light reconstruction

Given a stereo image pair, a neural network has been designed and trained to predict how structured light patterns should project on the scene surfaces of each image, separately. Specif-

ically, the network projects binary and gray code structured light patterns, which consist of vertical bars of white and black colour. To predict these patterns, the UNet [RFB15] network is trained as a per pixel binary classifier; 1 (white) or 0 (black). These predicted patterns can then be used to generate disparity or depth maps by performing a 2D cross-correlation over the epipolar lines. This reveals how the network has understood the 3D scene.

The input to the UNet is a pair of rectified stereo images (which do not contain any projection patterns), concatenated along the channel axis, $\mathrm{I}=[\mathrm{I}_l;\mathrm{I}_r]$, $\mathrm{I}_l,\mathrm{I}_r \in \mathbb{R}^{H\times W\times C}$. The network is tasked with predicting structured light projections in each input stereo image, where every pixel requires a unique code along a horizontal epipolar line. Y denotes the ground truth patterns and P the neural network output which is defined as $P=[P_l;P_r]$, $P_l,P_r \in \mathbb{R}^{H\times W\times t}, 0 \leq p \leq 1, y \in Y, p \in P$. The parameter t denotes the number of projected patterns and $t = 8$ for the virtual dataset and $t = 9$ for the medical dataset.

Losses: Two losses were chosen for structured light learning. Since the network is trained to produce a binary mask, binary cross entropy (bce) loss defined in Eq. (7.1) is used as the primary loss. To emphasise the pattern edges, an L2 loss is used on the horizontal derivatives as in Eq. (7.2). These derivatives are calculated by convolving over the images using a Prewitt operator mask, generating $D = Y * \begin{bmatrix}-1 & 0 & 1\end{bmatrix}$ and $\hat{D} = P * \begin{bmatrix}-1 & 0 & 1\end{bmatrix}$ which are the derivatives of the ground truth and predicted output, respectively with elements $d \in D$; $\hat{d} \in \hat{D}$. These two individual losses are summed together in Eq. (7.3) with a weighting gain of $\lambda_1 = 1/80$.

$$\mathcal{L}_{bce} = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot log(p) + (1-y_i)\cdot log(1-p_i) \tag{7.1}$$

$$\mathcal{L}_{L2} = \frac{1}{N}\sum_{i=1}^{N} ||d_i - \hat{d}_i|| \tag{7.2}$$

$$L_{sl} = \mathcal{L}_{bce} + \lambda_1 \mathcal{L}_{L2} \tag{7.3}$$

Extracting disparity from structured light: To generate the disparity map using the proposed structure light projection network, cross correlation is performed along the epipolar lines of a pair of rectified stereo images, fully connected along the pattern dimension. A patch, represented by W_ρ, where k is the reference index, of size $17 \times 17 \times t$ is taken in P_l, and cross correlation is performed over a range of pixels in P_r along the epipolar lines. More specifically, for every pixel on the left image, the cross correlation is estimated for pixels on the right image which are along the epipolar line at a distance within the maximum considered disparity for each dataset, individually. In our work, the maximum disparity has been set to 25% of the width of the image. The cross correlation on the right image will start from the same position as the examined patch on the left image, striding with a step s. The patch size was chosen after tuning.

$$cc(\rho, s) = \sum_{(u,v)\in W_\rho} P_l(u, v) \cdot P_r(u - s, v), \forall s \in [0..W/4] \tag{7.4}$$

$$disparity = \arg\max_s cc(\rho, s) \tag{7.5}$$

7.3.3 Direct disparity estimation with multi task leaning

For this task, we are exploring the benefits of training disparity and structured light jointly. Specifically, we investigate the relative effect of introducing the MTL framework, whether or not the MTL framework improves the disparity estimation performance produced using STL. The PSMNet [CC18b] architecture and training procedure were chosen as it is one of the more recognised architectures for disparity estimation and has recorded competitive performances in benchmark challenges. However, to note, any other architecture could have been used. In this work, the PSMNet architecture was modified in such a way that no added complexity or increased number of parameters is introduced for the disparity estimation path. Compartmentalising the design allows the structured light section to be detached when training is complete.

The PSMNet contains a stacked hourglass module to regress disparity. The modification which

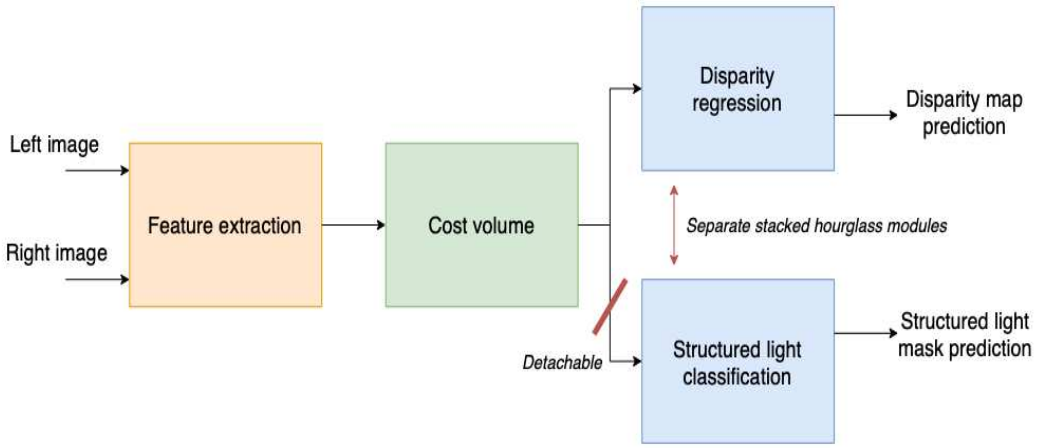


Figure 7.8: Simple block diagram representation of the multi task learning framework that has been applied on top of the PSMNet architecture.

we introduced to the architecture to create the multi task framework was done at this point. More specifically, the stacked hourglass module was duplicated; so that there is bifurcation after the cost volume. This results in two parallel paths, one for either task as shown in Figure 7.8. A disparity range of 96 was chosen. The training is performed on RGB images with dimensions 256×256. In the evaluation stage, the outputs are resized using bilinear interpolation to match the original dimensions which are 256×256 for the virtual images, 720×576 for the medical images and 1280×1024 for the SCARED images.

The loss function is composed of the structured light cross entropy loss Eq. (7.1) and an L2 disparity loss. In our work, three different weighting strategies have been used to combine these loss terms. This is because multi task learning is complex and different strategies benefit different tasks in a way which is generally unknown before experimentation is performed. Firstly, a simple constant scaling of the structured light loss was done with Eq. (7.6). Secondly, the training strategy from [LJD19] was implemented Eq. (7.7)-(7.8), which modifies the weights every epoch. Parameters λ_3 and λ_4 are the task weightings; a product of the ratio of the previous epoch's total losses. Here, $\zeta_1 = 2$ and $\zeta_2 = 2$. The parameter ζ_2 is used to balance the task weighting distribution, and ζ_1 is used to gain the softmax. The values were chosen after extensive experimentation. Thirdly, the uncertainty weighting strategy from [KGC18] was implemented as in Eq. (7.9). The weight $\lambda_5 = 0.5$ has been introduced to prioritise the disparity learning. The σ parameters denote the observed noise in practice, these are learnable

parameters used to weight each loss.

$$L_{const} = \lambda_2 \mathcal{L}_{sl} + \mathcal{L}_{disp} \tag{7.6}$$

$$\lambda_k = \zeta_1 \frac{exp(\omega_k \zeta_2)}{\sum_{i \in \{3,4\}} \omega_i \zeta_2}, \omega_k = \frac{\mathcal{L}_{task}(t-1)}{\mathcal{L}_{task}(t-2)}, k \in [3,4], task = \begin{cases} sl, & \text{if } k = 3 \\ disp, & \text{if } k = 4 \end{cases} \tag{7.7}$$

$$L_{epr} = \lambda_3 \mathcal{L}_{sl} + \lambda_4 \mathcal{L}_{disp} \tag{7.8}$$

$$L_{unc} = \frac{\lambda_5}{2\sigma_{sl}^2} \mathcal{L}_{sl} + \frac{1}{2\sigma_{disp}^2} \mathcal{L}_{disp} + log(\sigma_{sl}) + log(\sigma_{disp}) \tag{7.9}$$

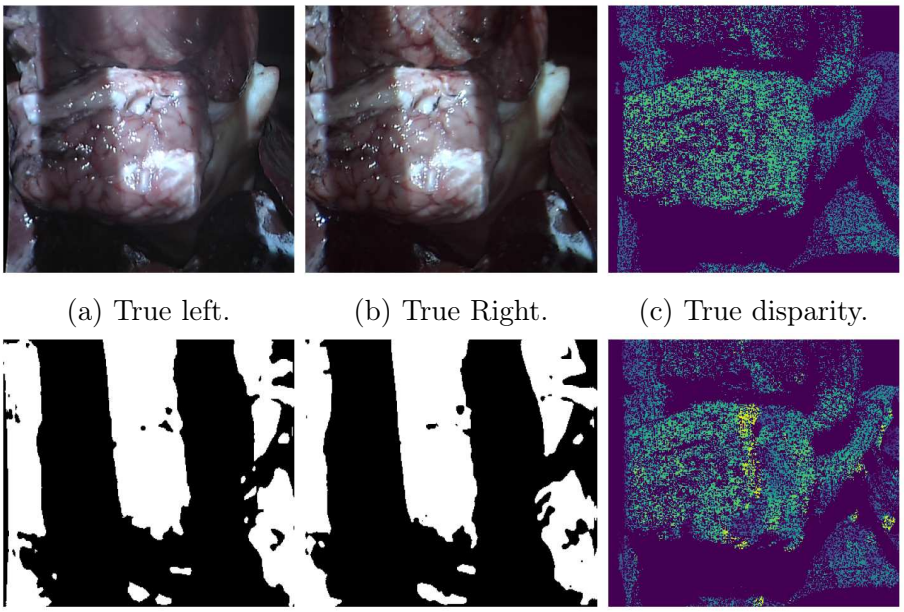
7.4 Results

7.4.1 Evaluation of structured light reconstruction

In this section, the reconstruction capability of the proposed artificially projected structured light model is assessed, using the disparity maps generated with cross correlation, as described in Sec. 7.3.2. The aim of this validation is not to compare the performance of the proposed structured light reconstruction model to state-of-the-art disparity estimation models. Rather, we want to explore the benefits of using this alternative approach for disparity estimation and evaluate whether or not the results from the proposed method meet the expected levels of accuracy produced by the conventional approach of directly regressing disparity. More specifically, the hypothesis that this validation aims to verify is that using the structured light in the training process, through Multi-Task Learning (MTL), offers improved performance compared to Single-Task Learning (STL). Without requiring extra parameters for the disparity estimation path and not requiring extra data collection; alleviating issues of data limitation.

A comparison network was trained to directly estimate disparity; using the same UNet architecture. The network was trained to predict disparity using a scaled Sigmoid. The output is mapped to the width of the image. An L2 on the disparities was used as the loss function. This network was used to control the experiments and remove the influence of the network architecture on the comparison of the results. There are no major architectural differences between the network proposed for structured light reconstruction and this comparison network. With the exception of the final layer which has been modified to accommodate the output dimension requirements. Comparison to state-of-the-art disparity estimation models is out of the scope of this validation as our focus is to prove the viability of reconstructing structure light.

The 3D reconstruction results are presented in Tab. 7.1. The first and second column contain the mean absolute error (MAE) for each model. Both training and evaluation is performed on grayscale images of dimension 256×256. The network was trained on a single NVIDIA GeForce RTX 3080 10GB graphics card. Adam [KB15] was chosen as the optimizer, with a fixed learning rate of 0.0001. The deep learning model was coded using the PyTorch framework [PGM+19]



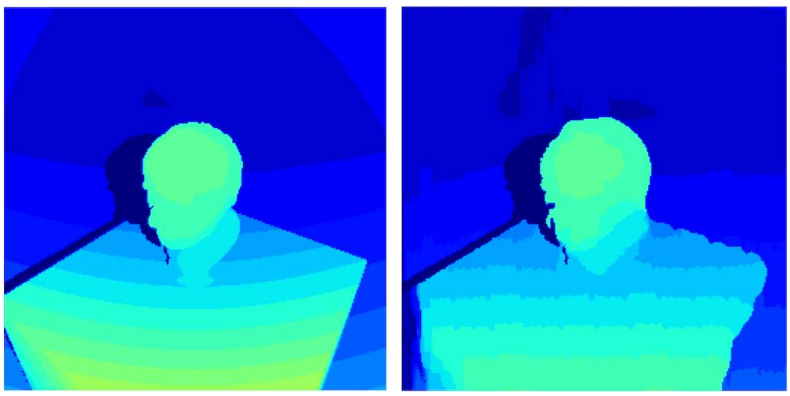
(a) True left. (b) True Right. (c) True disparity.

(d) Predicted left. (e) Predicted right. (f) Predicted disparity.

Figure 7.9: This figure contains an example output from the structured light network applied on the medical dataset. The top row contains the ground truth images and the bottom row shows the predictions.

Virtual dataset validation: The results on the virtual dataset, at the first two rows of Tab. 7.1, show that it is possible to achieve good reconstruction using the proposed approach - a sample output is shown in Figure 7.10. Virtual Seg denotes the metrics for the virtual object and the podium, segmented using masks provided by VisionBlender. Comparison between both approaches shows that comparable performance can be expected. For the virtual data, the training provided to the disparity network is higher quality than the structured light, as the disparity is extracted directly from the simulation environment. We hypothesise that this is the primary cause of the performance difference between the two implementations. The training/test split was 80/20, at the object level.

Medical dataset validation: Taking the networks trained on the virtual scenes, fine tuning on the medical dataset, shown in Figure 7.9, is performed to reveal the benefit of using structured light when the size of the datasets are limited. The training data for the structured light was 9 times larger than for the direct disparity network because 9 patterns were collected for



(a) Ground truth (b) Cross correlation output

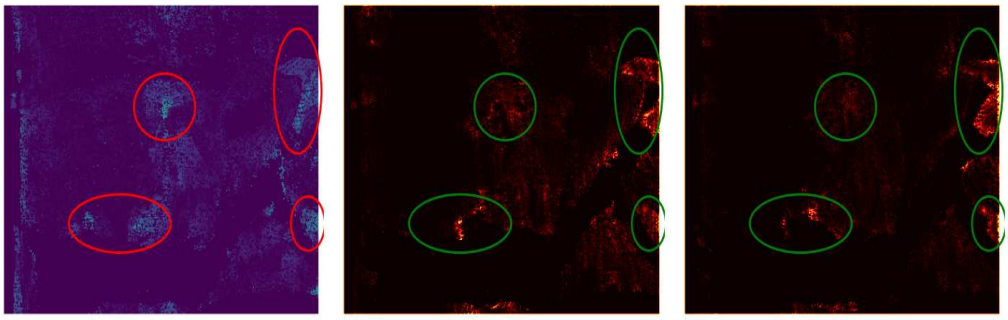
Figure 7.10: Sample from virtual dataset validation.

every depth map. The results are shown in the bottom row of Tab. 7.1. The accuracy recorded from the proposed method was twice as high as the direct regression approach. This verifies that having the larger volume of data for structured light training, when the datasets are small, improves the accuracy of the estimated disparity because of the increased complexity of the task and the increased number of training samples. The training/test split was 80/20, at the keyframe level.

Table 7.1: Mean absolute error (MAE) produced by the proposed structured light network (with the disparity produced using the cross correlation) and the direct disparity network.

Dataset	**Models** MAE (*pixels*)	
	Structured light	**Direct disparity**
Virtual	1.72	0.52
Virtual Seg	1.45	0.75
Medical	2.36	5.89

Uncertainty: As the network performs classification, it is easy to acquire the confidence in the predictions [CWS18]; the closer the Sigmoid output is to 0.5 the less confident the network is. An empirical assessment of the correlation between the disparity error and network confidence was made. It was concluded that there was a correlation between confidence and accuracy. An example of this is shown in Figure 7.11. The ellipses highlight the same areas in each image. These results show a correlation between areas of high uncertainty and areas of low accuracy. Figure 7.12 displays a histogram plot of the uncertainty and error in the above example. Each



(a) Disparity error. (b) Left uncertainty. (c) Right uncertainty.

Figure 7.11: Disparity accuracy and uncertainty correlation. The brighter the pixel, the greater the uncertainty and the less accurate. The ellipses are drawn over the same pixels in all three images. Correlating high energy is observed in the areas within the ellipses.

histogram bin represents the number of occurrences for each combination of uncertainty and disparity error. A positive correlation is seen between the disparity error and the uncertainty. This is numerical evidence of the correlation observed in Figure 7.11. Analysing confidence is critical for healthcare application and the adoption of deep learning methods in clinical practice. This is an advantage for using structured light reconstruction, over disparity estimation, as this is naturally a classification task (where the classification is the final output).

7.4.2 Evaluation of the benefit of using multi task learning for the enhancement of disparity learning

In this section, we explore the benefit of using joint training for the purpose of improving the performance of the disparity estimation compared to when performing single task learning. 5 algorithms are explored here. The standard PSMNet for single task learning (disparity) is used as the benchmark; denoted by *stl*. Then the modified, multi task learning PSMNet is trained 4 times using 4 different strategies. Firstly, L_{const} in Eq. (7.6) is implemented twice, where $\lambda_1 = 0.5, 10$; empirically chosen. For simplicity we denote the constant gain algorithms using: $cg_a := (\lambda_1 = 10)$ and $cg_b := (\lambda_1 = 0.5)$. Secondly, L_{epr} in Eq. (7.8) is implemented; denoted by *epr*. Finally, L_{unc} in Eq. (7.9) is implemented; denoted by *unc*.

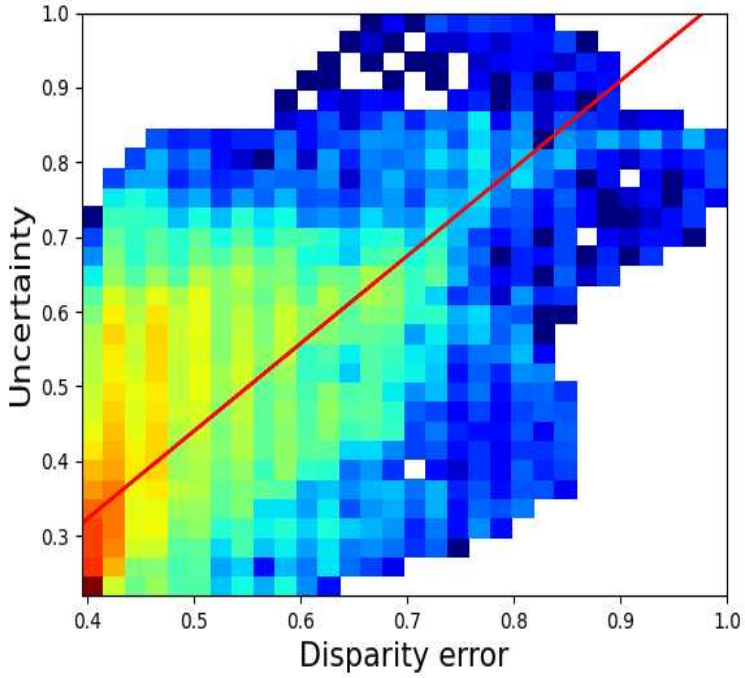


Figure 7.12: Log scaled correlation between uncertainty and prediction error. With a line of best fit drawn, in red, highlighting the positive correlation. The colourmap is set to 'jet'. The x and y axis are both normalised.

Virtual dataset validation: Firstly, all algorithms are assessed using the virtual dataset. The results are shown in Tab. 7.2. Again, Full is the MAE for the entire scene and Seg is the MAE for the podium and object only. This experiment is performed twice, firstly on the entire training set and secondly, on 1/16th of the training set; to see the impact of learning on a smaller dataset. 120 epochs were used for training; we determined this number after multiple experiments, balancing computation time and convergence. For the entire dataset experiment, all networks perform quite similarly, but with two of the multi task learning algorithms achieving slightly better results. However, on the 1/16th experiment the relative performance of *stl* drops, which highlights the regularizing benefits of using the proposed MTL framework. The extra complexity and the extra data has prevented overfitting to the fewer training samples. We observe that cg_b also drops in ranking; highlighting the difficulty of balancing the weighting distribution.

Table 7.2: Comparison of multi task learning vs single task learning on the virtual dataset. Two types of experiments were run, the first was using the full training set, the second was using 1/16th of the training set. Assessing the general performance of each algorithm when training on a large and a small dataset.

Rank	**Full training dataset**		**1/16th training dataset**	
	Full MAE (*pixels*)	**Seg** MAE (*pixels*)	**Full** MAE (*pixels*)	**Seg** MAE (*pixels*)
1st	cg_b : 0.203	cg_b : 0.380	cg_a : 0.496	cg_a : 0.745
2nd	unc : 0.207	unc : 0.381	cg_b : 0.504	unc : 0.778
3rd	**stl : 0.208**	**stl : 0.382**	unc : 0.516	epr : 0.781
4th	epr : 0.217	epr : 0.384	epr : 0.522	**stl : 0.790**
5th	cg_a : 0.217	cg_a : 0.389	**stl : 0.549**	cg_b : 0.808

Generalisability evaluation: We explore the generalisability properties of the networks trained on the entire virtual dataset by evaluating them on the SCARED dataset and the dense Middlebury 2014 Stereo training dataset. The virtual dataset that was used for training is limited, and therefore, benchmark results were not expected to be achieved. This test was constructed only to compare the performance of using STL and MTL; investigating the impact of the MTL framework with structured light learning. The SCARED dataset contained two test sets (TS8 and TS9), which collectively contain 45 unique images, which were warped to create a much larger volume of test data. So, to avoid the influence of error created during the artificial warping, only the 45 unmodified and complete images are used for evaluation. Tab. 7.3 shows the results of this experiment. The main point of interest here is the position of the *stl* algorithm. In all datasets, the *stl* algorithm performs worst. What can be understood from these results is that even though the performance of *stl* was comparable to the multi task learning performances on the virtual dataset, it has overfitted to the virtual dataset distribution. Whereas, the MTL framework has provided a regularizing effect during training, which has resulted in greater generalisability.

Medical dataset validation: Commonly in medical imaging, training data with ground truth is limited. To tackle this limitation, the standard strategy is to pre train a neural network on general and large datasets and then fine tune on data specific to the task. To replicate

Table 7.3: Generalisability evaluation using SCARED [AMW+21] and the Middlebury 2014 [SHK+14]. The choice of displaying the depth error for SCARED and disparity error for Middlebury reflects what is commonly done. The results show that the *stl* model has achieved poorer generalisability than the MTL approaches. TS denotes the test sets in SCARED. IQR denotes the interquartile range

Rank	**SCARED - TS 8**		**SCARED - TS 9**		**Middlebury**	
	Depth MAE (*mm*)	**Depth** IQR	**Depth** MAE (*mm*)	**Depth** IQR	**Disparity** MAE (*pixels*)	**Disparity** IQR
1st	unc : 25.15	unc : 21.71	unc : 27.86	unc : 24.63	unc : 58.41	unc : 76.17
2nd	cg_a : 29.01	cg_a : 27.94	cg_a : 32.41	cg_a : 28.10	epr : 61.43	epr : 76.20
3rd	cg_b : 29.96	cg_b : 31.36	epr : 31.92	cg_b : 31.07	cg_b : 63.27	cg_b : 81.48
4th	epr : 31.38	epr : 35.43	cg_b : 34.62	epr : 31.70	cg_a : 64.89	cg_a : 83.30
5th	**stl : 42.10**	**stl : 73.78**	**stl : 51.07**	**stl : 78.83**	**stl : 75.06**	**stl : 94.76**

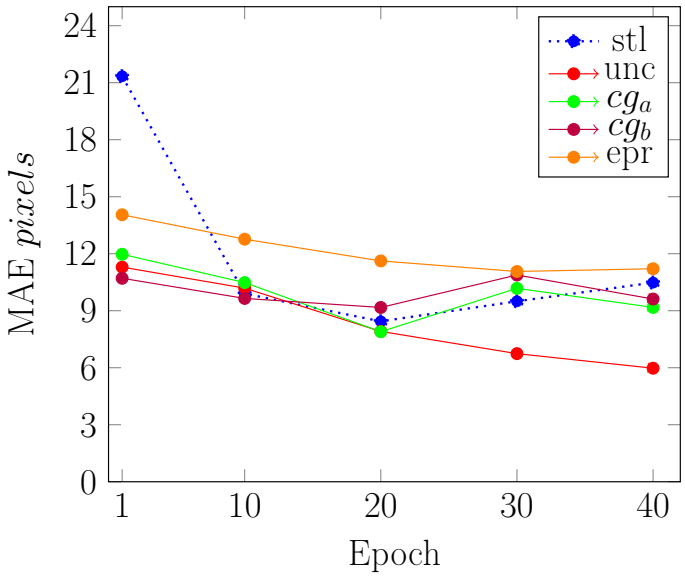


Figure 7.13: The blue dotted line is the performance of the *stl* model, which is used as a benchmark. What can be seen is that the only stably training algorithm is *unc*, in red, which also produces the best accuracy. All other models begin to overfit after 20 epochs.

this scenario, we use the models developed on the entire virtual dataset and fine tune on the medical dataset that we created. Each network was fine tuned for 40 epochs. The results shown in Figure 7.13 highlight the training stability of the compared models. The blue dotted line represents the *stl* performance. Overfitting can be inferred from the descent behaviour across the 40 epochs, due to the divergence of the validation accuracy. This is also reflected in the performance for the MTL models: *epr*, cg_a, cg_b. However, for *unc* the training is stable and it also achieves the best disparity MAE performance. Which again demonstrates the benefit of

the multi task learning framework, when the correct training strategy is implemented.

7.5 Conclusion

In this chapter, a novel approach to solving disparity estimation has been proposed that uniquely uses structured light information. We have proposed the first neural network to artificially project structured light patterns onto stereo images. This has allowed 3D reconstruction to be achieved using simple post processing similarity metrics. The proposed 3D reconstruction approach requires no explicit depth information during training. The performance evaluation results show that the proposed model accurately respects the surface geometry and achieves similar performance when compared to a direct regression network. As this proposed approach uses classification, it is also possible to estimate confidence in the disparity predictions, which is critical for tasks with high risk. This research was then extended, by designing a novel MTL framework to jointly predict structured light and disparity. Our validation verifies that introducing this MTL framework improves the generalisability and capability of learning from small datasets, for disparity estimation. All without increasing the number of parameters for the disparity estimation and using data which is already available and does not require extra annotation. Specifically, the MTL model *unc* produced results that were consistently better than *stl*, which demonstrates that when the correct multi task learning strategy is implemented, this is a better approach for developing a direct disparity estimation network. Our future work will focus on expanding our database to allow further validation of our work.

Chapter 8

Inexpensive Brain Model With Tumour Simulation

8.1 Introduction

Intraoperative ultrasound (IOUS) is an increasingly popular tool, as it is a viable alternative to more complex, time consuming and expensive solutions such as intraoperative magnetic resonance imaging [GB18] [KSS+14] However, who performs this task is not well defined, whether it should be the surgeon, radiologist, sonographer, or a combination of these professionals. The background commensurate training, in turn, remains highly variable and not standardised. In radiology, radiologists and sonographers undergo formal training and routinely perform hundreds of supervised ultrasound (US) scans on real patients over a condensed period to reach competency with accepted forms of assessment and accreditation. In contrast, in surgery, US training is highly variable depending on local practice and expertise and typically lacks the same structure. This is compounded by the much lower frequency of cases and greater risks that are inherent to intraoperative US which greatly limits access and slows experience in contrast to routine nonoperative diagnostic US which is high volume, low risk, and ubiquitous across most hospitals. Considering these challenges but also the many strengths of IOUS, such as real-time imaging and ease of integration into the surgical workflow, methods to improve

surgical training in IOUS are needed.

A particularly noteworthy application of IOUS is in oncological brain surgery, where the goal is maximal safe resection (removing the most tumour while preserving function and avoiding disability). The difficulty with gaining necessary IOUS experience is particularly daunting in neurosurgery, where it is uniquely not possible to perform practice US scanning of the brain on people outside of the operating theatre as a window through the skull is needed. This further steepens the learning curve impacting adoption and performance.

Traditionally, medical phantoms are used to simulate anatomy and medical procedures. There is limited access to purchasable educational tools and known homemade methods for creating phantoms for brain IOUS. Most commercially available brain phantoms are primarily designed for basic surgical training and tend to be expensive. As ultrasound is only one facet of the surgical process, the anatomical correctness of the tissue is usually simplistic, lacking real structural and textural details. This, in the case of the brain, is a significant issue due to the inherent complexity of the anatomy.

Homemade phantoms can be a more affordable and customisable solution. These phantoms can be made using food products [WNBL17] such as spam [NHCD16] or solutions such as ballistic gelatin [AKS+15] [MCB16], bovine gelatin [MJ15], sodium alginate [Aoy21], polyvinyl alcohol (PVA) [MSM+20] or agar [AEG+23] [MDP+15]. In particular, agar has been shown to successfully replicate the mechanical properties of human tumour tissue [SRDS17]. Although these methods are inexpensive, the core structure of the phantom will be simplistic and homogenous. Foreign bodies can be embedded in the medium to create diversity in echogenic features [SSSM20]. However, in combination this is not realistic. Real tissue is complex, and the interpretation of the ultrasound perspective and visible anatomy is one of the fundamental challenges in IOUS.

This chapter highlights the importance of creating anatomical accuracy in phantoms. A method is proposed to create a phantom to mimic tumours in inexpensive brain ex vivo using agar.

8.2 Materials and Methods

Table 8.1: Cost, acquisition and setup

Category	UpSurgeon	Animal Brain
Cost	High	Low
Availability	Moderate	Easy
Time to construct	5 min	10 min
Special storage	No	Yes
Special handling	No	Yes
Complexity of set up	Simple	Intermediate



Figure 8.1: UpSurgeOn brain box phantom synthetic material structure. The inner red material is the simulated tumour.

A summary of the cost and difficulty of creating the phantom is provided in Table 8.1. As a benchmark for commercially available phantoms, the 'Tumour box' and 5 'Glioma Cartridge' from UpSurgeOn [UpS24], were purchased at the cost of 654.92 € and 855.00 € with a shipping fee of 50.00 €. This phantom is designed primarily for resection practice, with ultrasound property being only a secondary feature. The homogeneity of the synthetic material is highlighted in Figure 8.1. However, this was the reason for the preference of this phantom, as with most other medical training phantoms, the phantom will be interacted with in such a way that there will be destruction of the material. For example, phantoms are commonly used for vascular access, epidurals, joint injections, or aspiration. Therefore, having multiple cartridges is a significant benefit of this phantom, unlike alternatives such as the Adult Brain phantoms by True Phantom Solutions. Although the latter may have better imaging feedback, it costs more ($1,738 Basic, $3,090 Standard and $4,250 Complex model), it is non-replaceable and can only be used in such a way as to preserve the integrity of the material. No cutting, modification, or rough handling.

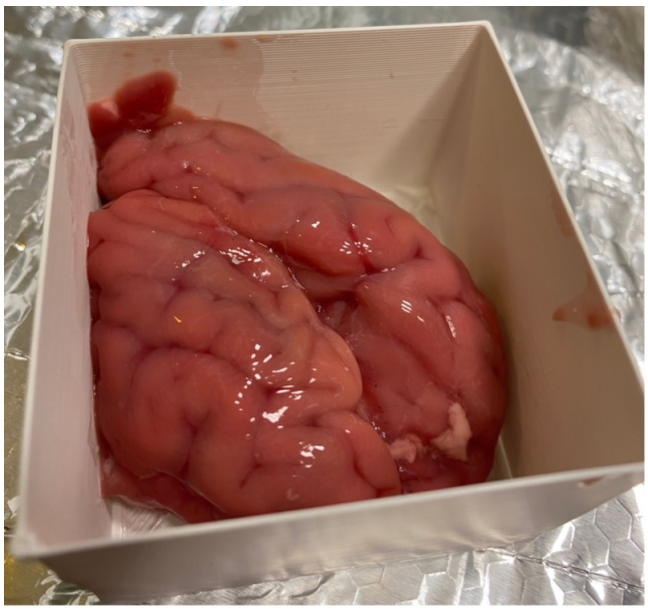


Figure 8.2: An ovine sample used to collect the data, before the agar solution was injected. The container has the dimensions 6.5cm x 7.0cm.

Multiple ovine and bovine ex vivo samples (Figure 8.2) were purchased at the cost of £5 each. All data were captured using an Ultrasonix Sonix CEP ultrasound system and a L14-5 linear probe. This was chosen as it is a previous generation machine, therefore producing lower quality images than can be expected with the modern state-of-the-art systems and mimicking a real-life scenario where clinicians rarely have access to the latest technological developments. Among the different materials used in the literature, to simulate brain tumours, agar was chosen due to its low cost and matching mechanical properties [14]. Agar is also easy to manipulate, since there is no need for freeze-thawing cycles (as needed with PVA [11]), and it can be injected during its liquid stage and allowed to solidify after injection in minimal time.

Tumour simulation was achieved using 2ml of 10-20% concentrate agar, heated using a Fisherbrand™ Hotplate Stirrer set to 150o and 1500 rpm, injected into the depth of the subcortical white matter/centrum semiovale - the agar was cooled before injection. Where both these areas are approximately equivalent to the human counterpart in terms of anatomical features such as texture, locations, colour, and physiological role. Depending on the concentration, temperature, volume, and area injected, from our observations, the time it took to harden could vary anywhere from a few minutes up to half an hour. To simulate the necrotic core of a tumour, water was injected into the hardened agar. Ultrasound feedback was used to guide the needle

into the agar solution.

Two experienced IOUS operators individually assessed the phantoms in relation to the acquired images. Evaluation categories were defined for qualitative assessment, using the five-point Liekert scale. Five categories were chosen, which were considered comprehensive for the evaluation of medical phantoms and their similarity to real brain tissue. The categories are Reusability, Echotexture, Cerebral structure, Tactile feedback, and Tumour mimic semblance.

8.3 Results

Table 8.2: Qualitative assessment using a five-point Liekert scale (1 = poor to 5 = excellent)

Feature	UpSurgeon	Animal Brain
Reusability	4	2
Echotexture	3	5
Cerebral structure	1	5
Tactile feedback	2	5
Tumour mimic semblance	2	4

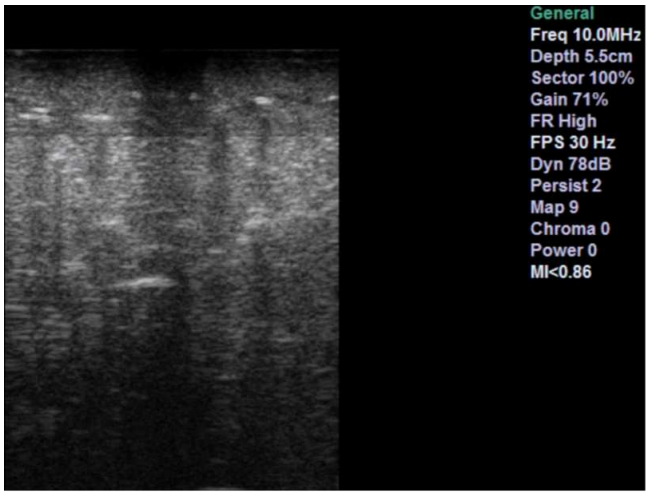


Figure 8.3: UpSurgeOn brain box phantom image from our machine.

The results of the qualitative ratings are shown in Table 8.2. Overall, the quality of the phantoms made using real brain tissue produced superior imaging, anatomical accuracy, and tumour accuracy, in comparison to the phantom. However, the animal phantom was not reusable and had to be disposed of on the day of use to prevent decay and necrosis.

8.3.1 Phantom analysis

Figure 8.3 shows an image of the phantom captured by our device. The most prominent and noticeable features are the linear, horizontal hyperechogenic lines. Although useful in defining the echogenic nature of the tissue analysed, their clinical correlation is debatable. The tumour itself appears as a circular hyperechoic ring with a hypoechoid core, causing significant acoustic shadowing. In the clinical setting, such shadowing is generally indicative of a tumour with

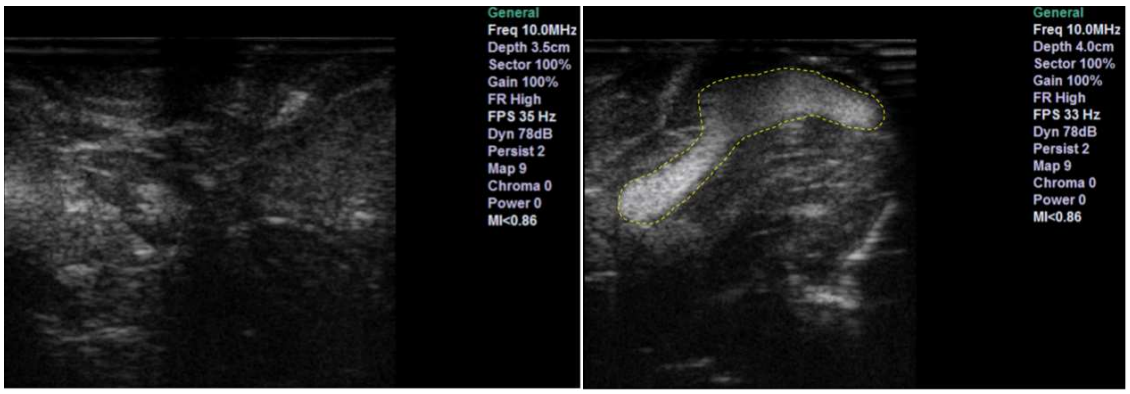


Figure 8.4: Bovine sample. Left shows ultrasound feedback of phantom without tumour addition. Right shows ultrasound feedback with agar injected.

a calcified component. In general, although there is echogenic diversity, these features are incoherent and do not resemble real anatomy.

Nonetheless, the reusability and absence of concern for necrosis and tissue degradation, makes the commercial phantom useful for basic US practice such as transducer manipulation, and contact force controlling.

8.3.2 Ex-Vivo and agar

The unmodified bovine brain sample is shown on the left side of Figure 8.4 and the ovine sample in section A and B of Figure 8.5. The good contrast within the tissue is the first noticeable property. The samples provided a good visualisation of the cortical sulci and gyri, with evident arachnoidal hyperechogenity and differentiation of grey-white matter. There were no shadows, artefacts, or elements of degradation. The images obtained through these specimens closely resembled a mammalian brain anatomy and therefore were considered more reliable to reproduce a clinical neurosurgical scenario.

In the same figures, the same tissues are also shown, but with the addition of injected 10-20% concentrate agar. In the context of the bovine brain, the addition of agar is shown in the right image of Figure 8.4. With the ovine tissue, sections C and D in Figure 8.5 show the agar solution. The agar solution was injected into the depth of the inferior frontal sulcus. As expected, the tumour appeared embedded in the subcortical white matter creating an echogenic

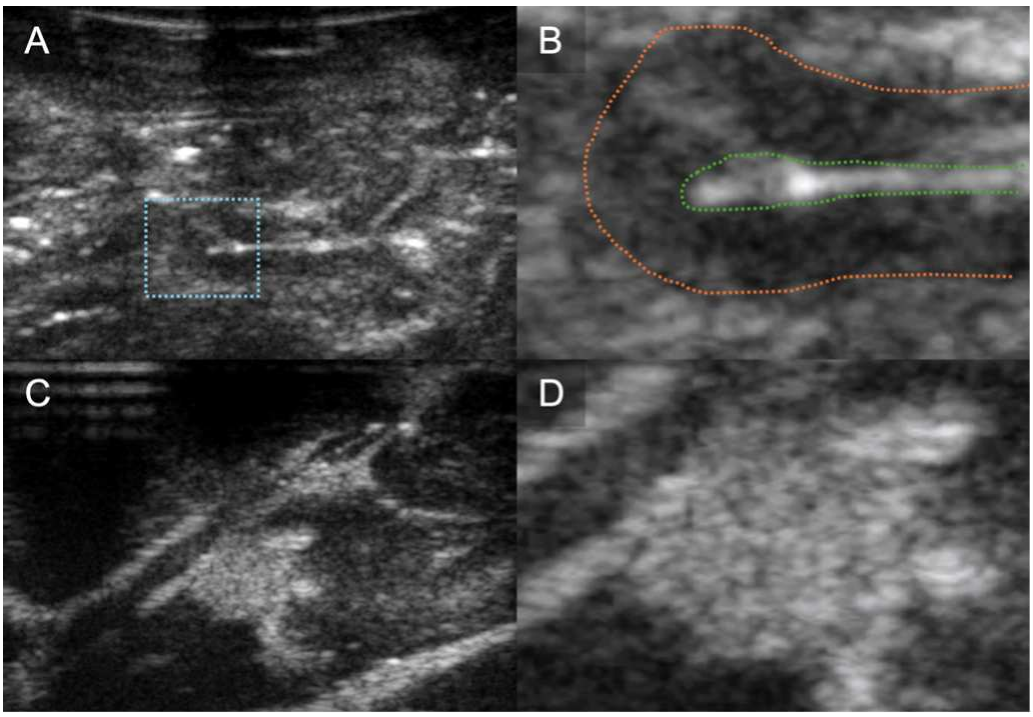


Figure 8.5: Ovine sample. B-mode image of sheep brain (A) with magnified image of gyrus (B) demonstrating discernible normal cerebral architecture with sulci and pia (outlined by green dotted line), cortex (outlined by orange dotted line) and surrounding white matter. B-mode image of the same sheep brain after localised infiltration of agar to mimic a hypercellular, echogenic tumour (C) with magnified image of tumour mimic demonstrating clear margins (D).

sphere, with a narrow tract following the entry point.

Unlike phantoms, agar injection into a real brain not only showed a well-defined and clearly identifiable anomaly but also provided a mass and an infiltrative effect to the surrounding areas. This interaction with brain tissue appears to be more reliable and useful for educational purposes and research applications.

For reference of a normal IOUS view of brain tumour tissue, Figure 8.6 shows images of glioblastomas - highlighted with yellow dotted lines - taken intraoperatively and right before resection, at Charing Cross Hospital, using a state-of-the-art system Canon i900 ultrasound machine [WDA+23a]. Noticeable are the layered tissue, which produces structured echogenic diversity. Similar to what is seen in the ovine and bovine tissue and dissimilar to what is seen in the phantom. The tumours share similar resemblance to that produced using agar, creating a region of concentrated, higher echogenicity with diffusion along the margins.



Figure 8.6: Human intraoperative ultrasound images of Glioblastoma. The tumours are highlighted using the yellow dotted lines. The left and right figures show a maximum depth of 5cm, while the middle figure shows a maximum depth of 4.5cm.

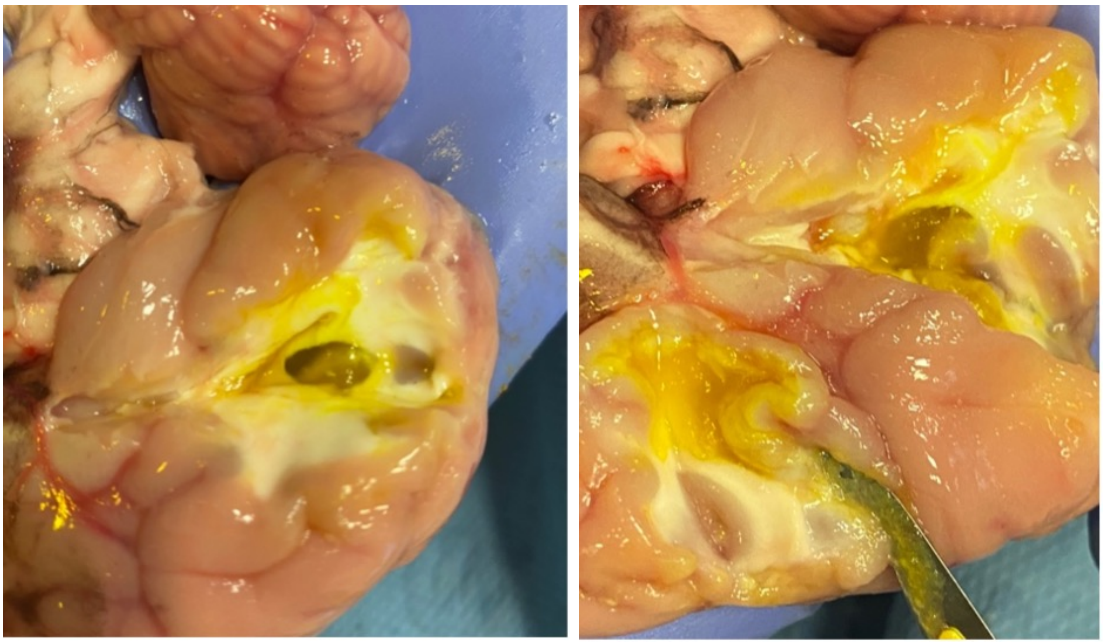


Figure 8.7: Cross section of one of the ovine brains. In this sample, two sites were injected with agar (dyed yellow)

A cross section of the agar in one of the ovine ex vivo samples is shown in Figure 8.8. The agar was mixed before injection with acriflavine to dye the tissue. As shown in the figure, it is possible to embed the agar solution in an area of normal tissue.

8.3.3 Necrosis Simulation

Figure 8.8 shows an example of the outcome of injecting water directly into the agar solution. In this example, the water successfully displaces the agar, creating a nested spherical area of

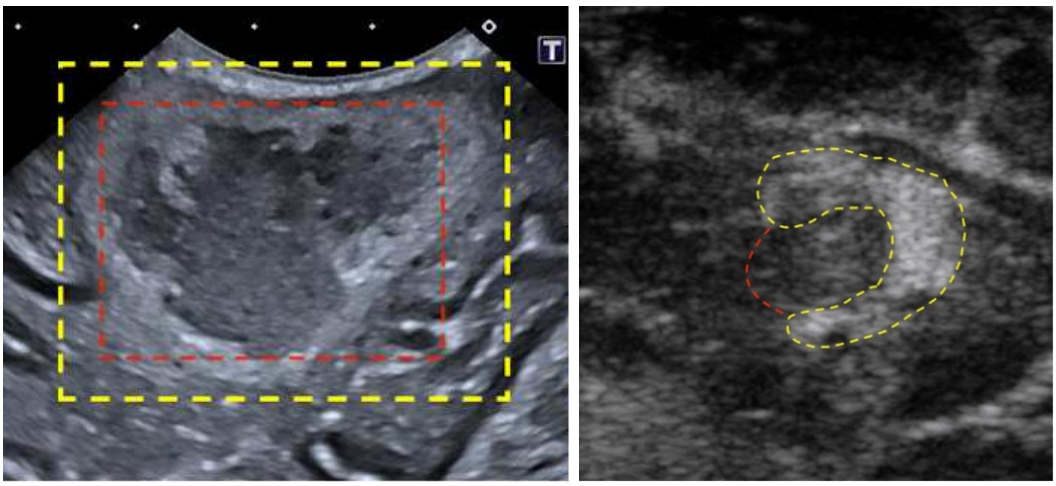


Figure 8.8: The right example shows a real case of a brain tumour, highlighted in yellow, with a necrotic core, highlighted in red. On the left is water added to the agar to simulate a necrotic core. Yellow line outlines the agar, red line outlines the water.

lower echogenicity within the higher echogenic agar solution. This can be seen as an area of slightly higher echogenicity than the background levels. The figure also shows an image of a human brain tumour with a necrotic core to highlight the impact on the echogenic profile.

8.4 Discussion

Table 8.3: Hypothetical training workshop itinerary utilising ex vivo models and agar solution

Session	**Overall Aim**	**Tasks**
Session 1	Basic US principles	Introduction to US for neurosonography; Recap of basic transducer manipulation; Demonstration and hands-on practice using synthetic phantom models.
Session 2	Introduction to ex vivo models	Presentation of animal brain ex vivo models; Application of basic principles using ex vivo models; Identification of anatomical landmarks/tissue characterisation.
Session 3	US Guided needling	Needle tracking using US; US guided needle injection of agar solution.
Session 4	Tumour Simulation	Localisation of prepared tumour simulations; US guided needle sampling from the simulated tumour; US guided incision to expose the tumour.

The creation of resources for surgical simulation, particularly the simulation of anatomy, is ever in demand and requires ongoing commitments to overcome a series of technical, logistical, and regulatory challenges. Regarding the specific problems related to the application of IOUS during brain tumour surgery, this chapter addresses one of the critical issues related to this imaging technology, the lack of simulation resources to support preclinical / cadaveric training and technical innovation.

Through our experiments with agar injection into animal ex vivo brain tissue, we show that it is possible to mimic tumours in inexpensive real tissue, replicating an anatomy that is as similar to human as possible and providing the expected ultrasound feedback. The simulation of necrosis, by injecting water into the hardened agar, created an acceptable resemblance to a high-grade real-life tumour scenario. This solution removed a few of the issues related to the artificial phantom, namely the lack of resemblance to the surrounding gyri and sulci, the acoustic shadow underlying the tumour model (absent in most cases) and overall, the signal obtained in terms of both resolution and reliability.

The creation of medical simulation resources such as this is important, as it provides a method of improving practical skills and reducing the dependency on learning from experience during real

surgeries. Based our experiences, we propose a hypothetical US workshop for neurosonography training, utilising ex vivo models as the primary resource. The workshop would begin with an introduction to basic US principals and progress towards more advanced techniques such as tumour localisation, transduce orientation for optimal imaging perspective, and US guided needle sampling. A detailed hypothetical workshop itinerary is outlined in Table 8.3. For further realism, the use of an animal head with craniotomy prepared would better simulate the surgical environment and should not be much harder to acquire than brains individually.

Extending from this, it would also facilitate learning new US-guided techniques such as US-guided core biopsies which are presently, due to lack of access, under applied in neurosurgery but readily used in other medical fields. For example, in this case the operators found that the task of using US to guide the agar injections was qualitatively associated with an improvement in US coordination and US localisation of the needle.

In addition, inexpensive, accurate simulation of brain anatomy and pathology encourages technical innovation. For example, data can be collected more easily (possibly to supplement the more difficult-to-acquire in vivo data), to enable deep learning training, which usually requires large volumes of data to prevent overfitting. Furthermore, the inexpensiveness and anatomical and textural accuracy of the phantom can provide a tool for conducting robotic experiments [JSN23]. As with most robotic applications for surgery, ultrasound guidance is still in an early preclinical, experimental research-focused phase. As such, this has allowed for temporary circumvention of the simulation of more challenging surgical procedures which may utilise ultrasound, and the expected intricacies of the anatomical features that would be present. In a large number of these robotic setups, the application is non-invasive, on-the-skin procedures [VLNZ21] [JZB+21], or are works that can be evaluated on simple synthetic medical phantoms, to explore technical novelty in the design and control of the robotic system [DWK+23] [ACS+21] [SHSM21] [FMP+23]. Nevertheless, evaluation of robotic systems on ex vivo is not uncommon. For example human cadavers for spine scanning [LDC+23], porcine tissue for needle tracking [JLC+24], kidney and heart for learning-from-demonstration [JBZ+23]. Germane to this chapter, [SGRwHK23] uses an unmodified chicken breast for the evaluation of focused US ablation. The study pertained to the feasibility of real-time detection of the ablation area and, as such,

the task did not require the ablated tissue to be pathological. However, the clinical motivation for this research and many others within the field of robotic ultrasound is therapeutic surgery. By using the method proposed in this chapter, or future similar methods, the tumour mimicking tissue can act as the target for the robotic experiment e.g. thermal ablation.

Our solution is not without limitations. First, prion disease represents a serious biohazard and needs to be addressed properly when manipulating the brains of dead animals. We have circumvented this problem opting for young animal brains (veals and lambs) specifically harvested for human consumption and have already gone through official clearance with the relevant regulatory authorities. If an adult animal brain is preferable for whatever reason, a much stricter protocol would be required. As an additional precaution measure to manage any biosafety risk, all biological material has been discarded according to standard safety protocols. Second challenge, working with ex vivo biological tissue should ideally be done in biosafe laboratories, unlike commercial phantoms, which could be used in any setting. This set-up is per se expensive and/or not always accessible for all centres. Finally, while the tissue does show much closer resemblance with the in vivo images, the possibility of degradation-related artefacts must be considered. For these experiments, tissue samples were kept out of the refrigerator at room temperature for more than an hour, and the harvesting time presumably took place 24 to 48 hours earlier. However, no significant necrosis or tissue degradation was observed within this time period, both on macroscopic inspection and on imaging. Some degree of manipulation-related damage was noticed especially at the level of the basal structures, but this was negligible for the purpose of our acquisitions. Some of the tissue that was left out was later frozen and used on another day, and it was observed that the fluctuation of temperature and the freezing process minimally impacted the usability of the tissue. In general, we found that this setting and time frame is acceptable, reliable, and replicable.

8.5 Conclusion

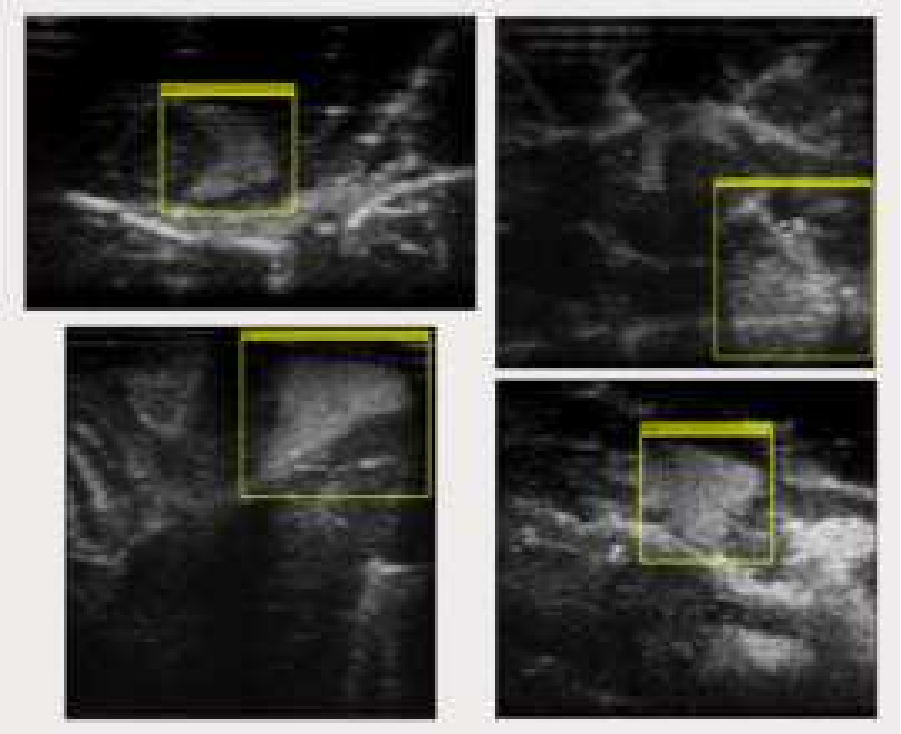


Figure 8.9: Samples from the dataset with super imposed bounding box annotation.

Methods for mimicking tumours within ultrasound images are valuable both for clinical training and research. This is especially the case with brain tumours, which are complex, highly variable, and diffuse. IOUS is a difficult tool to become proficient with, outside of real surgeries. The lack of affordable and accurate commercially available phantoms creates the need for home-made methods to create inexpensive homemade phantoms. Our proposed method solves this by mimicking tumours within real tissue, by combining easy and cheap butchers meat with agar powder. Furthermore, this will encourage technical innovation, as the lack of simulation resources has been a significant bottleneck in technical innovation. A dataset is created and made publicly available upon request that contains bounding-box annotations of multiple ex vivo ovine brains containing tumour-mimicking agar. The dataset contains 576 annotated images and 388 images without tumours, where the frames were extracted from US videos at 1Hz. The images are stored as MAT-files with the annotations in CSV files with the structure [top left corner, width, height]. Sample images from the dataset are shown in Figure 8.9.

Future work is encouraged to improve the realism of the tumour mimicry. This could include

testing different concentrations of agar solution, testing different locations, and using more than one compound to simulate tumour characteristics, creating more complex layering and diffusion effects. Further research is desirable to simulate a more infiltrative type of lesion, with reduced margin sharpness.

Chapter 9

Conclusion

In this thesis a broad spectrum of the topic of computer-assisted IOUS for brain tumour resection is explored. Treatment of brain tumours is difficult and the likelihood of mortality is high, even after successful treatment. Brain tumour resection is a uniquely challenging surgical task, where preoperative imaging offers minimal reliability during surgery and there is no binarity in the boundaries of the tumour. Intraoperative tools and techniques are variable across centres, where many avenues are being explored to improve the surgical process - however, these efforts are often limited by high costs and resource availability. As a consequence, minimal research output has been produced from both a clinical and a technical perspective. In particular, this has resulted in a lack of availability of data with annotation. This lack of output is more notable when considering the current developments in deep learning. Motivated by these factors, this thesis explores and proposes solutions to various translation challenges for IOUS, with an emphasis on data-drive and automated methods.

The main outcome from this work has been the formal identification and understanding of where the future challenges of computer vision for enabling IOUS are, and the proposal of solutions that solve key issues and help enable the development of robotic solutions. In particular, methods that tackle the eternal issue of data limitation, a method for analysing probe-tissue contact, as well as a study and methods deal with the issue of tumour margin uncertainty.

9.1 Contributions

9.1.1 Pathalogical Tissue Delineation

The first contribution is given in Chapter 3. The development of tumours and their structure within brain tissue is complex due to their diffuse and infiltrative properties. Precise delineation is achieved only through devices that can measure on a cellular or molecular level. However, these are impractical during surgery. IOUS is the most promising intraoperative tool; however, it lacks both the contrast between healthy and pathological tissue and the spatial resolution, which would need to be finer than the infiltrative margin and cellular spread, to clearly define the margins of pathological tissue. There is a lack of formal training and standardisation of IOUS, which has contributed to the sluggish adoption of the tool between centres. The study contained in the chapter was carried out to produce the first measure of the impact of uncertainty in the tumour margin and the impact of lack of formal training. A consultant neuroradiologist and three neurosurgeons, experienced with IOUS, were assigned to segment the tumours in a set of IOUS images captured during brain tumour resection. The primary experiment was the measure of the dissimilarity of neurosurgeons' annotations compared to the neuroradiologist, which was treated as the gold standard. An average intersection over union of 0.723 was measured, which shows variability in the annotations. This study formally addressed an issue that was only known informally among certain experienced users but is important to address if IOUS is to become a standardised tool.

The second contribution is given in Chapter 4. This chapter explores the topic of brain tumour detection in IOUS following the findings in the previous chapter. Transformers and attention based neural networks have shown considerable interest in recent years. However these models are known to be difficult to train on limited data. This chapter explores a newly proposed method of robust fine-tuning through constraining gradient descent to the pre-trained model. A dataset of 3705 images is annotated and 3 fold evaluation was performed, using the DETR and DEF DETR models trained in both unconstrained and constrained frameworks. This study explores methods to alleviate data limitations for tumour detection, where the results show that

training while leveraging the pre-trained parameters as an anchor, can produce superior IOU performances. Following the format unconstrained/constrained, the DETR IOU 0.475:0.577 and F1 0.571:0.716 and for DEF DETR IOU 0.581:0.590 and F1 0.720:0.756.

The third contribution is given in Chapter 5. As shown in the previous study, tumour segmentation is prone to interobserver variability due to factors including the inherent uncertainty caused by the lack of binarity of the margins. This chapter proposed a novel annotation method and protocol based on sparse demarcation using a confidence grading framework. The annotation was proved to successfully reflect the interobserver variance, with the confidence annotations achieving a brier score and soft dice similarity to the interobserver variance map of 0.00762 and 0.076. Further evaluation is performed to determine the usability of the annotations for neural network training. The confidence annotations and the segment annotations by the neuroradiologist were used to train a UNet and it was measured that the confidence annotation trained network outperformed the segmentation annotation for all test setups and across five-fold analysis - where the brier score was 0.02058 compared with 0.02466 and the soft dice similarity was 0.21974 compared with 0.29246. The proposed annotation method and protocol are a potential avenue for better annotation of brain tumours which better relects the infiltrative behaviour of brain tumours and can be a better annotation for data learning methods, as it is implicitly a better representation of the data than hard segments.

9.1.2 Probe Manipulation Analysis

The first contribution is given in Chapter 6. Control of the US probe during IOUS is determined primarily through the image feedback - not by looking at the probe. Therefore, automatic analysis of the quality of the probe-tissue coupling, for feedback on the positioning and orientation of the probe, is best through analysis of the image quality. In particular, determining whether there is visible tissue and the detection of contact-related artefacts. The method proposed is based on detection of points of dynamic change in the image - a type of edge detection but for nondefined edges. Through detection of dynamic points, a discrete representation of the visible tissue can be achieved through the construction of a simplicial complex. From this, a

contact-related acoustic shadow can be detected, as well as confidence in the visible tissue. The artefact detection performance was compared with an ablation, statistical, and a deep learning method. A three-fold analysis of the methods using intraoperative data was performed and the proposed method achieved the best F score, of 0.864, on the unseen data. The behaviour of the proposed method for constructing confidence maps was also evaluated against two established confidence mapping methods. Data was captured, where an US probe is orientated in and out of optimality (perpendicular to the surface), which creates an expected confidence trajectory in the scans. The three methods were evaluated, and the proposed method achieved the best performance in all scans with a normalised root mean square error of 0.168. Further experiments were performed using a robotic system to determine whether acoustic shadow detection was capable of guiding and optimising robotic scanning. The results showed that the acoustic shadow detection setup was capable of guiding and correcting the motion of multiple scans. The method proposed described in this chapter provides an automatic solution to the analysis of probe-tissue contact, which may facilitate automated scanning procedures or may be used for US trainee operators.

9.1.3 Resource Leveraging

Surgical data is limited and hard to acquire. In addition to this, simulation resources often lack anatomical realness and are expensive.

The first contribution is given in Chapter 7. Stereo vision is a popular method for reconstructing scenes due to the availability of high-resolution and frame rate cameras. Deep learning has demonstrated far better performance over nonlearning methods. Significant research has been conducted in this field. However, it is clear that there is a strong correlation between data volume and neural network performance when training using conventional supervised frameworks. Surgical data is limited, and there is no 3D craniotomy data available publicly, and acquiring ethics to capture data is difficult. To tackle this issue, a strategy is proposed to take advantage of structured light data, which is commonly captured and used to produce ground truth disparity for training, by training neural networks to learn projections in a multi-task learning framework

together with ground truth disparity. First, experimental evidence is provided, showing that it is possible to learn the projection of structured light in such a way that depth information is also maintained. In the medical dataset, the structured light trained network achieved superior performance in the test set than the disparity trained network, achieving 2.36($pixels$) mean absolute error in comparison to 5.89($pixels$). Then, it is demonstrated that training a state of the art neural network (PSMNet) in a multi task learning framework, for both structured light and disparity, improves the learning of disparity on the medical dataset. Within 40 epochs, the *unc* framework was the only training framework to achieve continuous convergence and also the best mean absolute error of 5.975($pixels$). Where the single task learning's best performance achieved 8.441($pixels$) and where training diverged after 20 epochs. The work in this chapter demonstrates the utilisation of commonly collected data to support learning for tasks that lack large datasets.

The second contribution is given in Chapter 8. In vivo experiments are an ethical minefield and would be impractical for early stage development. Medical phantoms have been shown to be an invaluable resource and an alternative to in vivo, for simulating anatomy and surgical procedures for technical development. Brain anatomy is complex and simulating it using synthetic materials is difficult. However, cheap commercially available animal brain ex vivo naturally has these features. This is highlighted by professional IOUS operators evaluating the US feedback from both a state-of-the-art synthetic phantom and ex vivo and graded each using a 5-point Likert scale assessment. The two key gradings for the image feedback were echotexture and cerebral structure. The synthetic phantom was graded with 3 and 1, while ex vivo with 5 and 4. The proposed method for mimicking tumour tissue was the injection of agar solution directly into ex vivo. The same professional IOUS users also graded the images of the synthetic tumour in both the synthetic phantom and the ex vivo, and graded the tumour mimic semblance as 2 and 4, respectively. This method provides researchers and clinical teachers with an inexpensive method for creating accurate mimicking of brain tumours in US.

9.2 Limitations and Future Work

This section identifies areas that were not explored and therefore limits some of the assumptions or conclusions that were made.

9.2.1 Craniotomy 3DV Study

A significant majority of 3DV output is entirely focused on beating benchmarks on popular datasets and exploring technical novelties. Although this is a necessary step to improve technical methodologies, in healthcare cases, these developments offer limited benefit in improving understanding of clinical challenges and do not provide evidence of translation prospects. The work in this thesis was explored under the assumption that fine-tuning would be required, from both a performance perspective and also an ethical perspective. Real surgical environments are highly dynamic and have many variables. Considering the current level of 3DV capabilities, it is clear that efforts should be made to measure how these methods perform on craniotomies. Including a wide spread of algorithmic methods across different sensors, such as stereo and monocular vision, structured light, and LIDAR. This study would provide real evidence of what does and does not work, and why or why not. This is vital to ensure that progress is made while keeping in mind the real, ultimate application.

9.2.2 Ultrasound Machine Settings

US image quality is highly variable between different machines, especially between the machine settings. It is the job of the operator to ensure that the settings are optimally tuned for visualisation of the region of interest. Intraoperative data capture often precedes any knowledge of what the data is actually going to be used for; therefore, it can be assumed that most IOUS data captured will tend to suffer bias in favour of good visualisation of, for example, the pathalogical tissue within the image. This is an issue that affects large proportions of US research. In the experimental work in this thesis, data diversity was pursued, so that not all

images had, for example, the depth setting optimally chosen for the pathological tissue extent. However, control of this aspect of the data selection process (machine settings) was weak. For methods that, for example, pursue robotic ultrasound, more research is needed to address the setting issue. Whether it be ways to automate the setting calibration process or settings-free imaging. This will be difficult as it would require working with the hardware of the machines and the raw RF data, both of which are not always available.

9.2.3 Multicentre studies

The work related to pathology delineation was carried out with professionals exclusively from the Imperial College London NHS Trust. Although this was sufficient to highlight the issue of interobserver variability and the issue of the uncertainty of the tumour margin. For a more complete analysis, it is necessary to involve clinicians from other centres and other countries.

Bibliography

[ACB+19] Zeynettin Akkus, Jason C. Cai, Arunnit Boonrod, Atefeh Zeinoddini, Alexander D. Weston, Kenneth A. Philbrick, and Bradley James Erickson. A survey of deep-learning applications in ultrasound: Artificial intelligence-powered ultrasound for improving clinical workflow. Journal of the American College of Radiology : JACR, 16 9 Pt B:1318–1328, 2019.

[ACH+20] Mohammad Ashraf, Nabeel Choudhary, Syed Shahzad Hussain, Usman Ahmad Kamboh, and Naveed Ashraf. Role of intraoperative computed tomography scanner in modern neurosurgery – an early experience. Surgical Neurology International, 11, 2020.

[ACM+16] Felix Arlt, Claire Chalopin, Andrea Müns, Jürgen Meixensberger, and Dirk Lindner. Intraoperative 3d contrast-enhanced ultrasound (ceus): a prospective study of 50 patients with brain tumours. Acta Neurochirurgica, 158:685–694, 2016.

[ACS+21] Mojtaba Akbari, Jay Carriere, Ron S. Sloboda, Tyler Meyer, Nawaid Usmani, Siraj Husain, and Mahdi Tavakoli. Robot-assisted breast ultrasound scanning using geometrical analysis of the seroma and image segmentation. 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages 3784–3791, 2021.

[ADGKF19] Walid S. Al-Dhabyani, Mohammed Mohammed Mohammed Gomaa, H Khaled, and Aly A. Fahmy. Dataset of breast ultrasound images. Data in Brief, 28, 2019.

[ADGKF20] Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. Dataset of breast ultrasound images. Data in Brief, 28:104863, 2020.

[AEG+23] Anastasia Antoniou, Nikolas Evripidou, Leonidas Georgiou, Antreas Chrysanthou, Cleanthis Ioannides, and Christakis Damianou. Tumor phantom model for mri-guided focused ultrasound ablation studies. Medical physics, 2023.

[AEP+22] Danilo Aleo, Ziad Elshaer, Andreas Pfnür, Patrick J. Schuler, Marco Maria Fontanella, Christian Rainer Wirtz, Andrej Paľa, and Jan Coburger. Evaluation of a navigated 3d ultrasound

integration for brain tumor surgery: First results of an ongoing prospective study. Current Oncology, 29:6594 – 6609, 2022.

[AKS+15] Richard Amini, Jeffrey Zane Kartchner, Lori A Stolz, David E. Biffar, Allan J. Hamilton, and Srikar Adhikari. A novel and inexpensive ballistic gel phantom for ultrasound training. World journal of emergency medicine, 6 3:225–8, 2015.

[AKS+23] Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J. Fleet. Synthetic data from diffusion models improves imagenet classification. ArXiv, abs/2304.08466, 2023.

[ALH+24] Giulio Anichini, Maria Leiloglou, Zepeng Hu, Kevin O'Neill, and Daniel Elson. Hyperspectral and multispectral imaging in neurosurgery: a systematic literature review and meta-analysis. European journal of surgical oncology : the journal of the European Society of Surgical Oncology and the British Association of Surgical Oncology, page 108293, 2024.

[AMS22] Hazrat Ali, Shafaq Murad, and Zubair Shah. Spot the fake lungs: Generating synthetic medical images using neural diffusion models. In Irish Conference on Artificial Intelligence and Cognitive Science, 2022.

[AMW+21] Max Allan, A. Jonathan McLeod, Cong Cong Wang, Jean-Claude Rosenthal, Zhenglei Hu, Niklas Gard, Peter Eisert, Ke Fu, Trevor Zeffiro, Wenyao Xia, Zhanshi Zhu, Huoling Luo, Fucang Jia, Xiran Zhang, Xiaohong Li, Lalith Sharan, Thomas Kurmann, Sebastian Schmid, Raphael Sznitman, Dimitris Psychogyios, Mahdi Azizian, Danail Stoyanov, Lena Maier-Hein, and Stefanie Speidel. Stereo correspondence and reconstruction of endoscopic data challenge. ArXiv, abs/2101.01133, 2021.

[ANAS17] Suheil Artul, William Nseir, Zaher Armaly, and Michalle Soudack. Superb microvascular imaging: Added value and novel applications. Journal of Clinical Imaging Science, 7, 2017.

[Aoy21] Minoru Aoyagi. Sodium alginate ultrasound phantom for medical education. Ultrasonic Imaging, 43:253 – 261, 2021.

[APH20] Ahmed Z. Alsinan, Vishal M. Patel, and Ilker Hacihaliloglu. Bone shadow segmentation from ultrasound data for orthopedic surgery using gan. International Journal of Computer Assisted Radiology and Surgery, 15:1477 – 1485, 2020.

[Axl95] Sheldon Axler. Linear algebra done right. Undergraduate Texts in Mathematics, 1995.

[BA19] James A. Balogun and Ayodeji Matthew Adebayo. Perception (of) and willingness to choose a neurosurgery career among final-year medical students in ibadan, nigeria. World neurosurgery, 2019.

[BB82] Dana H. Ballard and Christopher M. Brown. Computer vision. 1982.

[BBL+16] Timothy J. Brown, Matthew C Brennan, Michael Li, Ephraim W. Church, Nicholas J. Brandmeir, Kevin Rakszawski, Akshal S. Patel, Elias B Rizk, Dima Suki, Raymond Sawaya, and Michael J. Glantz. Association of the extent of resection with survival in glioblastoma: A systematic review and meta-analysis. JAMA oncology, 2 11:1460–1469, 2016.

[BCC+22] Bahareh Behboodi, François-Xavier Carton, Matthieu Chabanas, Sandrine de Ribaupierre, Ole Solheim, Bodil Karoline Ravn Munkvold, Hassan Rivaz, Yiming Xiao, and Ingerid Reinertsen. Resect-seg: Open access annotations of intra-operative brain tumor ultrasound images. ArXiv, abs/2207.07494, 2022.

[BCH+14] Benjamin Berkels, Ivan Cabrilo, Sven Haller, Martin Rumpf, and Karl Schaller. Co-registration of intra-operative brain surface photographs and pre-operative mr images. International Journal of Computer Assisted Radiology and Surgery, 9:387–400, 2014.

[BCML16] Florian Berton, Farida Cheriet, Marie-Claude Miron, and Catherine Laporte. Segmentation of the spinous process and its acoustic shadow in vertebral ultrasound images. Computers in biology and medicine, 72:201–11, 2016.

[BDK+21] Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. Explaining neural scaling laws. Proceedings of the National Academy of Sciences of the United States of America, 121, 2021.

[Bec19] J. Beck. Scan the world. https://www.myminifactory.com/scantheworld, 2019.

[Bel59] William A. Belson. Matching and prediction on the principle of biological classification. Journal of The Royal Statistical Society Series C-applied Statistics, 8:65–75, 1959.

[BGV92] Bernhard E. Boser, Isabelle M Guyon, and Vladimir Naumovich Vapnik. A training algorithm for optimal margin classifiers. In Annual Conference Computational Learning Theory, 1992.

[BH14] Alison Burgess and Kullervo Hynynen. Drug delivery across the blood–brain barrier using focused ultrasound. Expert Opinion on Drug Delivery, 11:711 – 721, 2014.

[BI21] Lorenzo Brigato and Luca Iocchi. A close look at deep learning with small data. 2020 25th International Conference on Pattern Recognition (ICPR), pages 2490–2497, 2021.

[Bis07] Christopher M. Bishop. Pattern Recognition and Machine Learning. Springer, 2007.

[BJD+24] Yuanwei Bi, Zhongliang Jiang, Felix Duelmer, Dianye Huang, and Nassir Navab. Machine learning in robotic ultrasound imaging: Challenges and perspectives. Annu. Rev. Control. Robotics Auton. Syst., 7, 2024.

[BJF+09] Eric Barua, Jeffrey S. Johnston, Jeffrey Fujii, Roger Dzwonczyk, Ennio Antonio Chiocca, and Sergio D. Bergese. Anesthesia for brain tumor resection using intraoperative magnetic resonance imaging (imri) with the polestar n-20 system: experience and challenges. Journal of clinical anesthesia, 21 5:371–6, 2009.

[BJT+21] Dhiego Chaves De Almeida Bastos, Parikshit Juvekar, Yanmei Tie, Nick Jowkar, Steve Pieper, Willam M Wells, Wenya Linda Bi, Alexandra Golby, Sarah Frisken, and Tina Kapur. Challenges and opportunities of intraoperative 3d ultrasound with neuronavigation in relation to intraoperative mri. Frontiers in oncology, 11:656519, 2021.

[BKM+16] Christian F. Baumgartner, Konstantinos Kamnitsas, Jacqueline Matthew, Tara P. Fletcher, Sandra Smith, Lisa M. Koch, Bernhard Kainz, and Daniel Rueckert. Sononet: Real-time detection and localisation of fetal standard scan planes in freehand ultrasound. Ieee Transactions on Medical Imaging, 36:2204 – 2215, 2016.

[BP66] Leonard E. Baum and Ted Petrie. Statistical inference for probabilistic functions of finite state markov chains. Annals of Mathematical Statistics, 37:1554–1563, 1966.

[BPK+18] Alexander V. Buslaev, Alex Parinov, Eugene Khvedchenya, Vladimir I. Iglovikov, and Alexandr A Kalinin. Albumentations: fast and flexible image augmentations. ArXiv, abs/1809.06839, 2018.

[Bra] Brain tumour research. Accessed on 20th January 2025.

[Bri50] Glenn W. Brier. Verification of forecasts expressed in terms of probability. Monthly Weather Review, 78:1–3, 1950.

[BWC+21a] Jannis Born, Nina Wiedemann, Manuel Cossio, Charlotte Buhre, Gabriel Brändle, Konstantin Leidermann, Avinash Aujayeb, Michael Moor, Bastian Rieck, and Karsten Borgwardt. Accelerating detection of lung pathologies with explainable ultrasound image analysis. Applied Sciences, 11(2):672, Jan 2021.

[BWC+21b] Jannis Born, Nina Wiedemann, Manuel Cossio, Charlotte Buhre, Gabriel Brändle, Konstantin Leidermann, Avinash Aujayeb, Michael Moor, Bastian Rieck, and Karsten Borgwardt. Ac-

celerating detection of lung pathologies with explainable ultrasound image analysis. Applied Sciences, 11(2):672, Jan 2021.

[CBE+18] Veronica Corona, Martin Benning, Matthias Joachim Ehrhardt, Lynn F. Gladden, Richard Mair, Andi Reci, Andrew J. Sederman, Stefanie Reichelt, and Carola-Bibiane Schoenlieb. Enhancing joint reconstruction and segmentation with non-convex bregman iteration. Inverse Problems, 35, 2018.

[CBM+19] Liang Chen, Paul Bentley, Kensaku Mori, Kazunari Misawa, Michitaka Fujiwara, and Daniel Rueckert. Self-supervised learning for medical image analysis using image context restoration. Medical image analysis, 58:101539, 2019.

[CC18a] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5410–5418, 2018.

[CC18b] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 5410–5418, 2018.

[CDT+05] Olivier Clatz, Hervé Delingette, Ion-Florin Talos, Alexandra J. Golby, Ron Kikinis, Ferenc A. Jolesz, Nicholas Ayache, and S. Warfield. Robust nonrigid registration to capture brain shift from intraoperative mri. IEEE Transactions on Medical Imaging, 24:1417–1427, 2005.

[Cer98] Paul E. Ceruzzi. A history of modern computing. 1998.

[CESR+25] Santiago Cepeda, Olga Esteban-Sinovas, Roberto Romero, Vikas Kumar Singh, Prakash Shetty, Aliasgar Moiyadi, Ilyess Zemmoura, Giuseppe Roberto Giammalva, Massimiliano Del Bene, Arianna Barbotti, Francesco DiMeco, Timothy Richard West, Brian Vala Nahed, Ignacio Arrese, Roberto Hornero, and Rosario Sarabia. Real-time brain tumor detection in intraoperative ultrasound using yolo11: From model training to deployment in the operating room. 2025.

[CESS+25] Santiago Cepeda, Olga Esteban-Sinovas, Vikas Kumar Singh, Prakash Shetty, Aliasgar Moiyadi, Luke Dixon, Alistair Weld, Giulio Anichini, Stamatia Giannarou, Sophie Camp, Ilyess Zemmoura, Giuseppe Roberto Giammalva, Massimiliano Del Bene, Arianna Barbotti, Francesco DiMeco, Timothy Richard West, Brian Vala Nahed, Roberto Romero, Ignacio Arrese, Roberto Hornero, and Rosario Sarabia. Deep learning-based glioma segmentation of 2d intraoperative ultrasound images: A multicenter study using the brain tumor intraoperative ultrasound database (bratious). Cancers, 17 2, 2025.

[CH67] Thomas M. Cover and Peter E. Hart. Nearest neighbor pattern classification. IEEE Trans. Inf. Theory, 13:21–27, 1967.

[Che10] Jie Cheng. Cuda by example: An introduction to general-purpose gpu programming. Scalable Comput. Pract. Exp., 11, 2010.

[CKN17] Pierre Chatelain, Alexandre Krupa, and Nassir Navab. Confidence-driven control of an ultrasound probe. IEEE Transactions on Robotics, 33:1410–1424, 2017.

[CLS+15] Junghwan Cho, Kyewook Lee, Ellie Shin, Garry Choy, and Synho Do. How much data is needed to train a medical image deep learning system to achieve necessary high accuracy. arXiv: Learning, 2015.

[CMS+20] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. ArXiv, abs/2005.12872, 2020.

[Com18] Blender Online Community. Blender - a 3D modelling and rendering package. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018.

[Com23] Competition and Markets Authority. AI Foundation Models: Initial Report, 2023.

[cru] Cancer research uk: Brain, other cns and intracranial tumours statistics. Accessed on 20th January 2025.

[CSD+22] Soumick Chatterjee, Alessandro Sciarra, Max Dünnwald, Pavan Tummala, Shubham Kumar Agrawal, Aishwarya Jauhari, Aman Kalra, Steffen Oeltze-Jafra, Oliver Speck, and A. Nürnberger. Strega: Unsupervised anomaly detection in brain mris using a compact context-encoding variational autoencoder. Computers in biology and medicine, 149:106093, 2022.

[CTL+20] João Cartucho, Samyakh Tukra, Yunpeng Li, Daniel S. Elson, and S. Giannarou. Visionblender: a tool to efficiently generate computer vision datasets for robotic surgery. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, 9:331 – 338, 2020.

[CVB08] Giorgio G. Carrabba, Lashmi Venkatraghavan, and Mark Bernstein. Day surgery awake craniotomy for removing brain tumours: Technical note describing a simple protocol. Minimally Invasive Neurosurgery, 51:208 – 210, 2008.

[CWH+21] João Cartucho, Chi Hui Wang, Baoru Huang, Daniel S. Elson, Ara Darzi, and Stamatia Giannarou. An enhanced marker pattern that achieves improved accuracy in surgical tool tracking. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, 10:400 – 408, 2021.

[CWS18] Yuanzhouhan Cao, Zifeng Wu, and Chunhua Shen. Estimating depth from monocular images as classification using deep fully convolutional residual networks. IEEE Transactions on Circuits and Systems for Video Technology, 28:3174–3182, 2018.

[CWT+23] João Cartucho, Alistair Weld, Samyakh Tukra, Haozheng Xu, Hiroki Matsuzaki, Taiyo Ishikawa, Min Sung Kwon, Yong Eun Jang, Kwang-Ju Kim, Gwang Lee, Bizhe Bai, Lueder A. Kahrs, Lars Boecking, Simeon Allmendinger, Leopold Muller, Yitong Zhang, Yueming Jin, Bano Sophia, Francisco Vasconcelos, Wolfgang Reiter, Jonas Hajek, Bruno Leonardo Barros Silva, Lukas R. Buschle, Estevão Lima, João L. Vilaça, Sandro Queirós, and Stamatia Giannarou. Surgt challenge: Benchmark of soft-tissue trackers for robotic surgery. Medical image analysis, 91:102985, 2023.

[CXW+23] Siman Cai, Hao Xing, Yuekun Wang, Yu Wang, Wenbin Ma, Yuxin Jiang, Jianchu Li, and Hongyan Wang. Intraoperative shear-wave elastography and superb microvascular imaging contribute to the glioma grading. Journal of Clinical Neuroscience, 110:92–99, 2023.

[CXW+24] Siman Cai, Hao Xing, Yuekun Wang, Yu Wang, Wenbin Ma, Yuxin Jiang, Jianchu Li, and Hongyan Wang. Clinical application of intraoperative ultrasound superb microvascular imaging in brain tumors resections: contributing to the achievement of total tumoral resection. BMC Medical Imaging, 24, 2024.

[CZZ+24] Sung-Liang Chen, Quan Zhou, E. Zhixuan Zeng, Ashkan Ebadi, Adrian Florea, and Alexander Wong. Covid-net l2c-ultra: An explainable linear-convex ultrasound augmentation learning framework to improve covid-19 assessment and monitoring. Sensors (Basel, Switzerland), 24, 2024.

[DBK+20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. ArXiv, abs/2010.11929, 2020.

[DCH24] Ana Davila, Jacinto Colan, and Yasuhisa Hasegawa. Comparison of fine-tuning strategies for transfer learning in medical image classification. Image and Vision Computing, 146, 2024.

[DCLT19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In North American Chapter of the Association for Computational Linguistics, 2019.

[DDPT18] Thomas Deffieux, Charlie Demené, Mathieu Pernot, and Mickael Tanter. Functional ultrasound

neuroimaging: a review of the preclinical and clinical state of the art. Current Opinion in Neurobiology, 50:128–135, 2018.

[DDS+09] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, K. Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition, pages 248–255, 2009.

[DeA01] L M DeAngelis. Brain tumors. N. Engl. J. Med., 344(2):114–123, January 2001.

[DES+23] Raman Dutt, Linus Ericsson, Pedro Sanchez, Sotirios A. Tsaftaris, and Timothy M. Hospedales. Parameter-efficient fine-tuning for medical image analysis: The missed opportunity. In International Conference on Medical Imaging with Deep Learning, 2023.

[DHM+20] Alexander D'Amour, Katherine A. Heller, Dan I. Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D. Hoffman, Farhad Hormozdiari, Neil Houlsby, Shaobo Hou, Ghassen Jerfel, Alan Karthikesalingam, Mario Lucic, Yi-An Ma, Cory Y. McLean, Diana Mincu, Akinori Mitani, Andrea Montanari, Zachary Nado, Vivek Natarajan, Christopher Nielson, Thomas F. Osborne, Rajiv Raman, Kim Ramasamy, Rory Sayres, Jessica Schrouff, Martin G. Seneviratne, Shannon Sequeira, Harini Suresh, Victor Veitch, Max Vladymyrov, Xuezhi Wang, Kellie Webster, Steve Yadlowsky, Taedong Yun, Xiaohua Zhai, and D. Sculley. Underspecification presents challenges for credibility in modern machine learning. ArXiv, abs/2011.03395, 2020.

[Dix23] Sheila Dixon. Diagnostic Imaging Dataset Annual Statistical Release 2022/23. NHS England, 2023.

[DLGS+22] Luke Dixon, Adrian Lim, Matthew Grech-Sollars, Dipankar Nandi, and Sophie J. Camp. Intraoperative ultrasound in brain tumor surgery: A review and implementation guide. Neurosurgical Review, 45:2503 – 2515, 2022.

[DLR18] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society: Series B (Methodological), 39(1):1–22, 12 2018.

[DR01] George J Dohrmann and Jonathan M Rubin. History of intraoperative ultrasound in neurosurgery. Neurosurgery Clinics of North America, 12(1):155–66, 2001.

[dSFSY+06] Esmeraldo dos Santos Filho, Yoshifumi Saijo, Tomoyuki Yambe, Akira Tanaka, and Makoto Yoshizawa. Segmentation of calcification regions in intravascular ultrasound images by adaptive thresholding. 19th IEEE Symposium on Computer-Based Medical Systems (CBMS'06), pages 446–454, 2006.

[DTH+24] Reuben Dorent, Erick Torio, Nazim Haouchine, Colin Galvin, Sarah F. Frisken, Alexandra J. Golby, Tina Kapur, and William M. Wells. Patient-specific real-time segmentation in trackerless brain ultrasound. ArXiv, abs/2405.09959, 2024.

[dVSG+24] Maria Correia de Verdier, Rachit Saluja, Louis Gagnon, Dominic Labella, Ujjwal Baid, Nour El-Hoda Mostafa Tahon, Martha Foltyn-Dumitru, Jikai Zhang, Maram M. Alafif, Saif Baig, Ken Chang, Gennaro D'Anna, Lisa Deptula, Diviya Gupta, Muhammad Ammar Haider, Ali Hussain, Michael Iv, Marinos Kontzialis, Paul Manning, Farzan Moodi, Teresa Nunes, Aaron B. Simon, Nico Sollmann, David Vu, Maruf Adewole, Jake Albrecht, Udunna C. Anazodo, Rongrong Chai, Verena Chung, Shahriar Faghani, Keyvan Farahani, Anahita Fathi Kazerooni, Eugenio Iglesias, Florian Kofler, Hongwei Bran Li, Marius George Linguraru, Bjoern H Menze, Ahmed W. Moawad, Yury Velichko, Benedikt Wiestler, Talissa Altes, Patil Basavasagar, Martin Bendszus, Gianluca Brugnara, Jaeyoung Cho, Yaseen Dhemesh, Brandon K.K. Fields, Filip Garrett, Jaime Gass, Lubomir Hadjiiski, Jona A. Hattangadi-Gluth, Christopher P Hess, Jessica L Houk, Edvin Isufi, Lester J. Layfield, George M. Mastorakos, John Mongan, Pierre Nedelec, Uyen Nguyen, Sebastian Oliva, Matthew W. Pease, Aditya Rastogi, Jason Sinclair, Robert Smith, Leo P. Sugrue, Jonathan Thacker, Igor Vidić, Javier E Villanueva-Meyer, Nate White, Mariam Aboian, Gian Marco Conte, Anders M. Dale, Mert Rory Sabuncu, Tyler Seibert, Brent D Weinberg, Aly H. Abayazeed, Raymond Y. Huang, Sevcan Turk, Andreas M. Rauschecker, Nikdokht Farid, Philipp Vollmuth, Ayman Nada, Spyridon Bakas, Evan Calabrese, and Jeffrey D. Rudie. The 2024 brain tumor segmentation (brats) challenge: Glioma segmentation on post-treatment mri. ArXiv, abs/2405.18368, 2024.

[DWK+23] Michael Dyck, Alistair Weld, Julian Klodmann, Alexander Kirst, Luke Dixon, G. Anichini, Sophie Camp, Alin Albu-Schaffer, and Stamatia Giannarou. Toward safe and collaborative robotic ultrasound tissue scanning in neurosurgery. IEEE Transactions on Medical Robotics and Bionics, 6:64–67, 2023.

[EAAAM22] Omar Elharrouss, Younes Akbari, Noor Almaadeed, and Somaya Ali Al-Maadeed. Backbones-review: Feature extraction networks for deep learning and deep reinforcement learning approaches. ArXiv, abs/2206.08016, 2022.

[EEBJ+19] Maarten Everts, Addison Elliott, Taylor Braun-Jones, Rob Court, Hans Johnson, Isaiah Norton, Josh Warner, Peter Fischer, Hans Meine, Ali Ghayoor, Tashrif Billah, Simon Ekström, Katrin Leinweber, JcNils, Jasper van de Gronde, and Dan Brown. mhe/pynrrd: v0.4.1 released. November 2019.

[Enc09] Yavor Enchev. Neuronavigation: geneology, reality, and prospects. Neurosurgical Focus, 27(3):E11, 2009.

[FAL17] Chelsea Finn, P. Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In International Conference on Machine Learning, 2017.

[FB81] Martin A. Fischler and Robert C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM, 24:381–395, 1981.

[FBKC+12] Andriy Fedorov, Reinhard Beichel, Jayashree Kalpathy-Cramer, Julien Finet, Jean-Christophe Fillion-Robin, Sonia Pujol, Christian Bauer, Dominique Jennings, Fiona Fennessy, Milan Sonka, John Buatti, Stephen Aylward, James V Miller, Steve Pieper, and Ron Kikinis. 3d slicer as an image computing platform for the quantitative imaging network. Magn. Reson. Imaging, 30:1323–1341, 2012.

[FCG+21] RÃ©mi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, AurÃ©lie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, LÃ©o Gautheron, Nathalie T.H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. Pot: Python optimal transport. Journal of Machine Learning Research, 22(78):1–8, 2021.

[FGW+18] Huan Fu, Mingming Gong, C. Wang, K. Batmanghelich, and D. Tao. Deep ordinal regression network for monocular depth estimation. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 2002–2011, 2018.

[FH89] Evelyn Fix and Joseph L. Hodges. Discriminatory analysis - nonparametric discrimination: Consistency properties. International Statistical Review, 57:238, 1989.

[FLRP+18] Jeffrey De Fauw, Joseph R. Ledsam, Bernardino Romera-Paredes, Stanislav Nikolov, Nenad Tomasev, Sam Blackwell, Harry Askham, Xavier Glorot, Brendan O'Donoghue, Daniela Visentin, George Van Den Driessche, Balaji Lakshminarayanan, Clemens Meyer, Faith Mackinder, Simon Bouton, Kareem Ayoub, Reena Chopra, Dominic King, Alan Karthikesalingam, Cian O. Hughes, Rosalind Raine, J. Weston Hughes, Dawn A. Sim, Catherine A. Egan, Adnan Tufail, Hugh Montgomery, Demis Hassabis, Geraint E. Rees, Trevor Back, Peng Tee Khaw, Mustafa Suleyman, Julien Cornebise, Pearse Andrew Keane, and Olaf Ronneberger. Clinically applicable deep learning for diagnosis and referral in retinal disease. Nature Medicine, 24:1342 – 1350, 2018.

[FMP+23] Giovanni Faoro, Sabina Maglio, Stefano Pane, Veronica Iacovacci, and Arianna Menciassi. An artificial intelligence-aided robotic platform for ultrasound-guided transcarotid revascularization. IEEE Robotics and Automation Letters, 8:2349–2356, 2023.

[FNB+15] Les R. Folio, Chelsye J Nelson, Menashe Benjamin, Ayelet Ran, Guy Engelhard, and David A. Bluemke. Quantitative radiology reporting in oncology: Survey of oncologists and radiologists. AJR. American journal of roentgenology, 205 3:W233–43, 2015.

[FS41] Lillian Feinstein and Martin Schwarzschild. Automatic integration of linear second-order differential equations by means of punched card machines. Review of Scientific Instruments, 12:405–408, 1941.

[FWN50] Lyle A French, John J Wild, and Donald Neal. Detection of cerebral tumors by ultrasonic pulses. pilot studies on postmortem material. Cancer, 3(4):705–708, 1950.

[GAB17] Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Unsupervised monocular depth estimation with left-right consistency. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 6602–6611, 2017.

[GAD21] Dana Gu, Kirill F. Andreev, and Matthew E. Dupre. Major trends in population growth around the world. China CDC Weekly, 3:604 – 613, 2021.

[GAVN11] Mohit Gupta, Amit Agrawal, Ashok Veeraraghavan, and Srinivasa G. Narasimhan. Structured light 3d scanning in the presence of global illumination. In CVPR 2011, pages 713–720, 2011.

[GB] Philippe Guerin and Ruma Bhargava. Poverty and infectious disease come together far too often — the solution could be data-driven.

[GB18] Rajashree Uday Gandhe and Chinmaya Pradeep Bhave. Intraoperative magnetic resonance imaging for neurosurgery – an anaesthesiologist's challenge. Indian Journal of Anaesthesia, 62:411 – 417, 2018.

[GBC16] I. Goodfellow, Y. Bengio, and A. Courville. Deep Learning. Adaptive Computation and Machine Learning series. MIT Press, 2016.

[GKOH+21] Ian J. Gerard, Marta Kersten-Oertel, Jeffery A. Hall, Denis Sirhan, and D. Louis Collins. Brain shift in neuronavigation of brain tumors: An updated review of intra-operative ultrasound applications. Frontiers in Oncology, 10, 2021.

[GLCA20] Charley Gros, Andréanne Lemay, and Julien Cohen-Adad. Softseg: Advantages of soft versus binary training for image segmentation. Medical image analysis, 71:102038, 2020.

[Gow75] John C. Gower. Generalized procrustes analysis. Psychometrika, 40:33–51, 1975.

[GPAM+14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Advances in neural information processing systems, pages 2672–2680, 2014.

[GRU10] Andreas Geiger, Martin Roser, and R. Urtasun. Efficient large-scale stereo matching. In ACCV, 2010.

[GS23] Mauro Giuffré and Dennis L. Shung. Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. NPJ Digital Medicine, 6, 2023.

[GSY13] Stamatia Giannarou, Marco Visentini Scarzanella, and Guang-Zhong Yang. Probabilistic tracking of affine-invariant anisotropic regions. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35:130–143, 2013.

[GVJ+22] Simon Graham, Quoc Dang Vu, Mostafa Jahanifar, Fayyaz Minhas, David R. J. Snead, and Nasir M. Rajpoot. One model is all you need: Multi-task learning enables simultaneous histology image segmentation and classification. Medical image analysis, 83:102685, 2022.

[HCG20] Alex Ling Yu Hung, Wanwen Chen, and John Michael Galeotti. Ultrasound confidence maps of intensity and structure based on directed acyclic graphs and artifact models. 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), pages 697–701, 2020.

[HCMC10] Pierre Hellier, Pierrick Coupé, Xavier Morandi, and D. Louis Collins. An automatic geometrical and statistical method to detect acoustic shadows in intraoperative ultrasound brain images. Medical image analysis, 14(2):195–204, 2010.

[HDIT97] Nobuhiko Hata, Takeyoshi Dohi, Hiroshi Iseki, and Kintomo Takakura. Development of a frameless and armless stereotactic neuronavigation system with ultrasonographic registration. Neurosurgery, 41(3):608–13, 1997.

[HDK11] Ciaran Scott Hill, Lalani Dias, and N. D. Kitchen. Perceptions of neurosurgery: a survey of medical students and foundation doctors. British Journal of Neurosurgery, 25:261 – 267, 2011.

[Hei18] Mattias P. Heinrich. Intra-operative ultrasound to mri fusion with a public multimodal discrete registration tool. In POCUS/BIVPCS/CuRIOUS/CPM@MICCAI, 2018.

[HFP+23] Ragnhild Holden Helland, Alexandros Ferles, André Pedersen, Ivar Kommers, Hilko Ardon, Frederik Barkhof, Lorenzo Bello, Mitchel S. Berger, Tora Dunås, Marco Conti Nibali, Julia Furtner, Shawn L. Hervey-Jumper, Albert J. S. Idema, Barbara Kiesel, Rishi Nandoe Tewarie,

Emmanuel Mandonnet, Domenique M. J. Müller, Pierre A. Robe, Marco Rossi, Lisa Millgård Sagberg, Tommaso Sciortino, Tom T.A. Aalders, Michiel Wagemakers, Georg Widhalm, Marnix G. Witte, Aeilko H. Zwinderman, Paulina Majewska, Asgeir Store Jakola, Ole Solheim, Philip C. de Witt Hamer, Ingerid Reinertsen, Roelant S. Eijgelaar, and David Bouget. Segmentation of glioblastomas in early post-operative multi-modal mri with deep neural networks. Scientific Reports, 13, 2023.

[HGL+24] Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. Parameter-efficient fine-tuning for large models: A comprehensive survey. Trans. Mach. Learn. Res., 2024, 2024.

[HJA20a] Jonathan Ho, Ajay Jain, and P. Abbeel. Denoising diffusion probabilistic models. ArXiv, abs/2006.11239, 2020.

[HJA20b] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS '20, 2020.

[HJB16] Shawn L. Hervey-Jumper and Mitchel S. Berger. Maximizing safe resection of low- and high-grade glioma. Journal of Neuro-Oncology, 130:269–282, 2016.

[HLB+19] Ouwen Huang, Will Long, Nick Bottenus, Gregg E. Trahey, Sina Farsiu, and Mark L. Palmeri. Mimicknet, matching clinical post-processing under realistic black-box constraints. 2019 IEEE International Ultrasonics Symposium (IUS), pages 1145–1151, 2019.

[HLK+15] Hampus Holmer, Adam Lantz, Teena Kunjumen, Samuel R. G. Finlayson, Marguerite M. Hoyler, Amani Siyam, Hernán Montenegro, Edward T Kelley, James Campbell, Meena Nathan Cherian, and Lars Hagander. Global distribution of surgeons, anaesthesiologists, and obstetricians. The Lancet. Global health, 3 Suppl 2:S9–11, 2015.

[HMWL+21] Andrew M. Hersh, Smruti Mahapatra, Carly Weber-Levine, Tolulope Awosika, John N. Theodore, Hesham Mostafa Zakaria, Ann Liu, Timothy F Witham, and Nicholas Theodore. Augmented reality in spine surgery: A narrative review. HSS Journal®: The Musculoskeletal Journal of Hospital for Special Surgery, 17:351 – 358, 2021.

[HNA+17] Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Frederick Diamos, Heewoo Jun, Hassan Kianinejad, Md. Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically. ArXiv, abs/1712.00409, 2017.

[Hog85] Neville Hogan. Impedance control: An approach to manipulation: Part I—Theory. Journal of Dynamic Systems Measurement and Control-transactions of The Asme, 1985.

[Hop82] John J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. Proceedings of the National Academy of Sciences of the United States of America, 79 8, 1982.

[HPJ+23] Shih-Cheng Huang, Anuj Pareek, Malte E. K. Jensen, Matthew P. Lungren, Serena Yeung, and Akshay S. Chaudhari. Self-supervised learning for medical image classification: a systematic review and implementation guidelines. NPJ Digital Medicine, 6, 2023.

[HPP+20] Johannes Hofmanninger, F. Prayer, Jeanny Pan, Sebastian Röhrich, Helmut Prosch, and Georg Langs. Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem. European Radiology Experimental, 4, 2020.

[HRG+17] Lifeng He, Xiwei Ren, Qihang Gao, Xiao Zhao, Bin Yao, and Yuyan Chao. The connected-component labeling problem: A review of state-of-the-art algorithms. Pattern Recognit., 70:25–43, 2017.

[HRRM18] Daniel A. Hashimoto, Guy Rosman, Daniela Rus, and Ozanan R. Meireles. Artificial intelligence in surgery: Promises and perils. Annals of Surgery, 268:70–76, 2018.

[HS97] Sepp Hochreiter and Jurgen Schmidhuber. Long short-term memory. Neural Computation, 9:1735–1780, 1997.

[HS03] Stephen J. Hentschel and Raymond E. Sawaya. Optimizing outcomes with maximal surgical resection of malignant gliomas. Cancer control : journal of the Moffitt Cancer Center, 10 2:109–14, 2003.

[HZ04] Richard Hartley and Andrew Zisserman. Multiple View Geometry in Computer Vision. Cambridge University Press, 2004.

[HZRS15] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 770–778, 2015.

[II23] Irena M. Ilic and Milena D. Ilic. International patterns and trends in the brain cancer incidence and mortality: An observational study based on the global burden of disease. Heliyon, 9, 2023.

[IJK+20] Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods, 18:203 – 211, 2020.

[INB+25] Javed Iqbal, Ansa Naseem, Muhammad Arsalan Bashir, Kivanc Yangi, Ismail Bozkurt, and Bipin Chaurasia. Global perspective of neurosurgery practice in lower middle-income countries: challenges, opportunities, and the path forward. Annals of Medicine and Surgery, 87:2532 – 2536, 2025.

[ION+17] Mami Ishikawa, Yasushi Ota, Mutsumi Nagai, Gen Kusaka, Yuichi Tanaka, and Heiji Naritaka. Ultrasonography monitoring with superb microvascular imaging technique in brain tumor surgery. World neurosurgery, 97:749.e11–749.e20, 2017.

[Jai18] Anil K. Jain. Fundamentals of digital image processing. Control of Color Imaging Systems, 2018.

[JBZ+23] Zhongliang Jiang, Yuanwei Bi, Mingchuan Zhou, Ying Hu, Michael Burke, , and Nassir Navab. Intelligent robotic sonographer: Mutual information-based disentangled reward learning from few demonstrations. The International Journal of Robotics Research, 43:981 – 1002, 2023.

[JDK+23] Parikshit Juvekar, Reuben Dorent, Fryderyk Kögl, Erickson Torio, Colton Barr, Laura Rigolo, Colin Galvin, Nick Jowkar, Anees Kazi, Nazim Haouchine, Harneet Cheema, Nassir Navab, Steve Pieper, William M Wells, Wenya Linda Bi, Alexandra Golby, Sarah Frisken, and Tina Kapur. The brain resection multimodal imaging database (ReMIND), 2023.

[JGZ+20] Zhongliang Jiang, Matthias Grimm, Mingchuan Zhou, Javier Esteban, Walter Simson, Guillaume Zahnd, and Nassir Navab. Automatic normal positioning of robotic ultrasound probe based only on confidence map optimization and force measurement. IEEE Robotics and Automation Letters, 5:1342–1349, 2020.

[JKA+22] Kiran Jabeen, Muhammad Attique Khan, Majed Alhaisoni, Usman Tariq, Yudong Zhang, Ameer Hamza, Arturas Mickus, and Robertas Damasevicius. Breast cancer classification from ultrasound images using probability-based optimal deep learning feature fusion. Sensors (Basel, Switzerland), 22, 2022.

[JLC+24] Zhongliang Jiang, Xuesong Li, Xiangyu Chu, Angelos Karlas, Yuanwei Bi, Yingsheng Cheng, K. W. Samuel Au, and Nassir Navab. Needle segmentation using gan: Restoring thin instrument visibility in robotic ultrasound. IEEE Transactions on Instrumentation and Measurement, 73:1–11, 2024.

[JRHP12] Songbai Ji, David W Roberts, Alex Hartov, and Keith D Paulsen. Intraoperative patient registration using volumetric true 3d ultrasound without fiducials. Medical physics, 39(12):7540–7552, 2012.

[JRP19] S. Jayanthi, H. Ranganathan, and M. Palanivelan. Segmenting brain tumour regions with fuzzy integrated active contours. IETE Journal of Research, 68:514 – 525, 2019.

[JSN23] Zhongliang Jiang, Septimiu E. Salcudeanb, and Nassir Navab. Robotic ultrasound imaging: State-of-the-art and future perspectives. Medical image analysis, 89:102878, 2023.

[JWA17] Sebastian Hoppe Nesgaard Jensen, Jakob Wilm, and Henrik Aanæs. An error analysis of structured light scanning of biological tissue. In Image Analysis, pages 135–145, Germany, 2017. Springer. Scandinavian Conference on Image Analysis, SCIA ; Conference date: 12-06-2017 Through 14-06-2017.

[JZB+21] Zhongliang Jiang, Yue Zhou, Yuanfei Bi, Mingchuan Zhou, Thomas Wendler, and Nassir Navab. Deformation-aware robotic 3d ultrasound. IEEE Robotics and Automation Letters, 6:7675–7682, 2021.

[KB15] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. CoRR, abs/1412.6980, 2015.

[KCD+14] Peter Kazanzides, Zihan Chen, Anton Deguet, Gregory S Fischer, Russell H Taylor, and Simon P DiMaio. An open-source research kit for the da vinci® surgical system. In 2014 IEEE international conference on robotics and automation (ICRA), pages 6434–6439. IEEE, 2014.

[KDBF+24] Singh Mathuria Kaushal-Deep, Abdul Rashid Bhat, Mohsin Fayaz, Gianluca Scalia, Irfan Robbani, Muhammed Afzal Wani, Mehershree Lodhi, and Bipin Chaurasia. Intraoperative variations in intra-axial brain tumor size after craniotomy: a prospective study with histopathological and tumor tissue composition correlation, introducing concepts of tumor expansion and tumor surfacing. Neurosurgical review, 48 1:9, 2024.

[KGC18] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 7482–7491, 2018.

[Kha87] Oussama Khatib. A unified approach for motion and force control of robot manipulators: The operational space formulation. JRA, 3(1):43–53, 1987.

[KHC+13] Nicole M. A. Krekel, Max H. Haloua, Alexander M F Lopes Cardozo, Roos H de Wit, Anne Marie Bosch, Louise M de Widt-Levert, Sandra Muller, Henk van der Veen, Elisabeth Bergers, Elly S. M. de Lange – de Klerk, Sybren L. Meijer, and M. Petrousjka van den Tol. Intraoperative ultrasound guidance for palpable breast cancer excision (cobalt trial): a multicentre, randomised controlled trial. The Lancet. Oncology, 14 1:48–54, 2013.

[KJH+21] Rabindra Khadka, Debesh Jha, Steven Hicks, Vajira Lasantha Thambawita, M. Riegler, Sharib Ali, and P. Halvorsen. Meta-learning with implicit gradients in a few-shot setting for medical image segmentation. Computers in biology and medicine, 143:105227, 2021.

[KMR+23] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross B. Girshick. Segment anything. ArXiv, abs/2304.02643, 2023.

[KS19] Davood Karimi and Septimiu E. Salcudean. Reducing the hausdorff distance in medical image segmentation with convolutional neural networks. IEEE Transactions on Medical Imaging, 39:499–513, 2019.

[KSH12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. Communications of the ACM, 60:84 – 90, 2012.

[KSS+14] Pieter Leonard Kubben, Felix Scholtes, Olaf Schijns, Mariel P ter Laak-Poort, Onno P.M. Teernstra, Alfons G. H. Kessels, Jacobus J. van Overbeeke, Didier Martin, and Henk van Santbrink. Intraoperative magnetic resonance imaging versus standard neuronavigation for the neurosurgical treatment of glioblastoma: A randomized controlled trial. Surgical Neurology International, 5, 2014.

[KW13] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. CoRR, abs/1312.6114, 2013.

[KWA+23] Ira Ktena, Olivia Wiles, Isabela Albuquerque, Sylvestre-Alvise Rebuffi, Ryutaro Tanno, Abhijit Guha Roy, Shekoofeh Azizi, Danielle Belgrave, Pushmeet Kohli, Alan Karthikesalingam, Ali Taylan Cemgil, and Sven Gowal. Generative models improve fairness of medical classifiers under distribution shifts. Nature Medicine, 30:1166 – 1173, 2023.

[KWKN12a] Athanasios Karamalis, Wolfgang Wein, Tassilo Klein, and Nassir Navab. Ultrasound confidence maps using random walks. Medical image analysis, 16 6:1101–12, 2012.

[KWKN12b] Athanasios Karamalis, Wolfgang Wein, Tassilo Klein, and Nassir Navab. Ultrasound confidence maps using random walks. Medical image analysis, 16 6:1101–12, 2012.

[LBD+89] Yann LeCun, Bernhard E. Boser, John S. Denker, Donnie Henderson, Richard E. Howard, Wayne E. Hubbard, and Lawrence D. Jackel. Handwritten digit recognition with a back-propagation network. In Neural Information Processing Systems, 1989.

[LBK+15] Valeria De Luca, Tobias Benz, Satoshi Kondo, Lars König, D. Lübke, Sven Rothlübbers, Oudom Somphone, Stéphane Allaire, Muyinatu A. Lediju Bell, Daniel Chung, Daniel Chung,

Amalia Cifor, Cristian Grozea, Matthias Günther, Jürgen W. Jenne, Till Kipshagen, Markus Kowarschik, Markus Kowarschik, Nassir Navab, Jan Ruhaak, Julia Schwaab, and Christine Tanner. The 2014 liver ultrasound tracking benchmark. Physics in Medicine and Biology, 60:5571 – 5599, 2015.

[LCO18] Théo Lacombe, Marco Cuturi, and Steve Oudot. Large scale computation of means and clusters for persistence diagrams using optimal transport. ArXiv, abs/1805.08331, 2018.

[LCS+09] Carlos J. Ledezma, Wei Chen, Victor Sai, Bonnie Freitas, Timothy F. Cloughesy, Johannes Czernin, and Whitney B. Pope. 18f-fdopa pet/mri fusion in patients with primary/recurrent gliomas: initial experience. European journal of radiology, 71 2:242–8, 2009.

[LDC+23] Ruixuan Li, Ayoob Davoodi, Yuyu Cai, Kenan Niu, Gianni Borghesan, Nicola Alessandro Cavalcanti, Aidana Massalimova, Fabio Carrillo, Christoph Johannes Laux, Mazda Farshad, Philipp Furnstahl, and Emmanuel B. Vander Poorten. Robot-assisted ultrasound reconstruction for spine surgery: from bench-top to pre-clinical study. International Journal of Computer Assisted Radiology and Surgery, 18:1613–1623, 2023.

[LGK+21] Meghan G. Lubner, Lori Mankowski Gettle, David H. Kim, Timothy J. Ziemlewicz, Nirvikar Dahiya, and Perry J. Pickhardt. Diagnostic and procedural intraoperative ultrasound: technique, tips and tricks for optimizing results. The British journal of radiology, page 20201406, 2021.

[Lin70] Seppo Linnainmaa. The representation of the cumulative rounding error of an algorithm as a taylor expansion of the local rounding errors. Master's Thesis, University of Helsinki, 1970.

[LJBB20] Hamid Laga, Laurent Valentin Jospin, Farid Boussaïd, and Bennamoun. A survey on deep learning techniques for stereo-based depth estimation. IEEE transactions on pattern analysis and machine intelligence, PP, 2020.

[LJD19] Shikun Liu, Edward Johns, and Andrew J Davison. End-to-end multi-task learning with attention. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 1871–1880, 2019.

[LKHK18] Ju Hwan Lee, Ga Young Kim, Yoo Na Hwang, and Sung Min Kim. Automatic detection of dense calcium and acoustic shadow in intravascular ultrasound images by dual-threshold-based segmentation approach. Sensors and Materials, 2018.

[LKO+03] Frank Lindseth, Jon Harald Kaspersen, Steinar Ommedal, Thomas Langø, Jon Bang, Jørn Hokland, Geirmund Unsgaard, and Toril A Nagelhus Hemes. Multimodal image fusion in

ultrasound-based neuronavigation: improving overview and interpretation by integrating preoperative mri with intraoperative 3d ultrasound. Computer Aided Surgery, 8(2):49–69, 2003.

[Llo82] Stuart P. Lloyd. Least squares quantization in pcm. IEEE Trans. Inf. Theory, 28:129–136, 1982.

[LMB+14] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In European Conference on Computer Vision, 2014.

[Low99] David G. Lowe. Object recognition from local scale-invariant features. Proceedings of the Seventh IEEE International Conference on Computer Vision, 2:1150–1157 vol.2, 1999.

[LPBM19] Jose Lamarca, Shaifali Parashar, Adrien Bartoli, and José M. M. Montiel. Defslam: Tracking and mapping of deforming scenes from monocular sequences. IEEE Transactions on Robotics, 37:291–303, 2019.

[LPW+21] David N. Louis, Arie Perry, Pieter Wesseling, Daniel J. Brat, Ian A. Cree, Dominique Figarella-Branger, Cynthia Hawkins, Ho-keung Ng, Stefan M. Pfister, Guido Reifenberger, Riccardo Soffietti, Andreas von Deimling, and David W. Ellison. The 2021 who classification of tumors of the central nervous system: a summary. Neuro-oncology, 2021.

[LSC+23] Dhairya A. Lakhani, D. S. Sabsevitz, Kaisorn L. Chaichana, Alfredo Quinones-Hinojosa, and Erik H. Middlebrooks. Current state of functional mri in the presurgical planning of brain tumors. Radiology. Imaging cancer, 5 6:e230078, 2023.

[LSP+19] Sarah Leclerc, Erik Smistad, João Pedrosa, Andreas Østvik, Frederic Cervenansky, Florian Espinosa, Torvald Espeland, Erik Andreas Rye Berg, Pierre-Marc Jodoin, Thomas Grenier, Carole Lartizien, Jan D'hooge, Lasse Lovstakken, and Olivier Bernard. Deep learning for segmentation using an open large-scale dataset in 2d echocardiography. IEEE Transactions on Medical Imaging, 38(9):2198–2210, 2019.

[LT93] E R Laws and K Thapar. Brain tumors. CA Cancer J. Clin., 43(5):263–271, September 1993.

[LTD21] Lahav Lipson, Zachary Teed, and Jia Deng. Raft-stereo: Multilevel recurrent field transforms for stereo matching. 2021 International Conference on 3D Vision (3DV), pages 218–227, 2021.

[LTR+06] D Lindner, C Trantakis, C Renner, S Arnold, A Schmitgen, J Schneider, and J Meixensberger. Application of intraoperative 3d ultrasound during navigated tumor resection. min-Minimally Invasive Neurosurgery, 49(04):197–202, 2006.

[LWLF+22] Njål Lura, Kari S. Wagner-Larsen, David Forsse, Jone Trovik, Mari K. Halle, Bjørn I. Bertelsen, Øyvind O. Salvesen, Kathrine Woie, Camilla Krakstad, and Ingfrid Salvesen Haldorsen. What mri-based tumor size measurement is best for predicting long-term survival in uterine cervical cancer? Insights into Imaging, 13, 2022.

[LXLZ22] Zhiying Lu, Hongtao Xie, Chuanbin Liu, and Yongdong Zhang. Bridging the gap between vision transformers and convolutional neural networks on small datasets. ArXiv, abs/2210.05958, 2022.

[MAA+22] Aashna Mehta, Wireko Andrew Awuah, Abdullahi Tunde Aborode, Jyi Cheng Ng, Katherine Candelario, Ines Margarida Pinto Vieira, H. Bulut, Abdul-Rahman Toufik, Mohammad Mehedi Hasan, and Vladyslav Sikora. Telesurgery's potential role in improving surgical access in africa. Annals of Medicine and Surgery, 82, 2022.

[MASC22] Vikram Melapudi, Chandankumar Aladahalli, K. S. Shriram, and Hsi-Ming Chang. Exploiting acquisition path to detect shadows in ultrasound images. In Medical Imaging, 2022.

[MBGY14] Clément Maria, Jean-Daniel Boissonnat, Marc Glisse, and Mariette Yvinec. The gudhi library: Simplicial complexes and persistent homology. In International Congress on Mathematical Software, 2014.

[MCB16] Dustin Morrow, Julia Cupp, and Joshua Seth Broder. Versatile, reusable, and inexpensive ultrasound phantom procedural trainers. Journal of Ultrasound in Medicine, 35, 2016.

[MCE+11] Mina S. Makary, Ennio Antonio Chiocca, Natali Erminy, María A. Antor, Sergio D. Bergese, Mahmoud Abdel-Rasoul, Soledad A. Fernandez, and Roger Dzwonczyk. Clinical and economic outcomes of low-field intraoperative mri-guided tumor resection neurosurgery. Journal of Magnetic Resonance Imaging, 34, 2011.

[MCH+16] Nader Mahmoud, Iñigo Cirauqui, Alexandre Hostettler, Christophe Doignon, Luc Soler, Jacques Marescaux, and José M. M. Montiel. Orbslam-based endoscope tracking and 3d reconstruction. In CARE@MICCAI, 2016.

[MDP+15] Georgios Menikou, Tetiana Dadakova, Matt Pavlina, Michael Bock, and Christakis Damianou. Mri compatible head phantom for ultrasound surgery. Ultrasonics, 57:144–52, 2015.

[MGB10] John M K Mislow, Alexandra J. Golby, and Peter McL. Black. Origins of intraoperative mri. Magnetic resonance imaging clinics of North America, 18(1):1–10, 2010.

[MHM+19] Qingjie Meng, James Housden, Jacqueline Matthew, Daniel Rueckert, Julia Anne Schnabel, Bernhard Kainz, Matthew Sinclair, Veronika A. M. Zimmer, Benjamin Hou, Martin Rajchl,

Nicolas Toussaint, Ozan Oktay, Jo Schlemper, and Alberto Gomez. Weakly supervised estimation of shadow confidence maps in fetal ultrasound imaging. IEEE Transactions on Medical Imaging, 38:2755–2767, 2019.

[MJ15] José Mario Martínez and Boguslaw J. Jarosz. 3d perfused brain phantom for interstitial ultrasound thermal therapy and imaging: design, construction and characterization. Physics in Medicine & Biology, 60:1879 – 1900, 2015.

[MJB+15] Bjoern H Menze, András Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin S. Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, Levente Lanczi, Elizabeth R. Gerstner, Marc-André Weber, Tal Arbel, Brian B. Avants, Nicholas Ayache, Patricia Buendia, D. Louis Collins, Nicolas Cordier, Jason J. Corso, Antonio Criminisi, Tilak Das, Hervé Delingette, Çağatay Demiralp, Christopher R. Durst, Michel Dojat, Senan Doyle, Joana Festa, Florence Forbes, Ezequiel Geremia, Ben Glocker, Polina Golland, Xiaotao Guo, Andaç Hamamci, Khan M. Iftekharuddin, Rajesh Jena, Nigel M. John, Ender Konukoglu, Danial Lashkari, José Antonio Mariz, Raphael Meier, Sérgio Pereira, Doina Precup, Stephen John Price, Tammy Riklin-Raviv, Syed M. S. Reza, Michael T. Ryan, Duygu Sarikaya, Lawrence H. Schwartz, Hoo-Chang Shin, Jamie Shotton, Carlos Alberto Silva, Nuno J. Sousa, Nagesh K. Subbanna, Gábor Székely, Thomas J. Taylor, Owen M. Thomas, N. Tustison, Gözde B. Ünal, Flor Vasseur, Max Wintermark, Dong Hye Ye, Liang Zhao, Binsheng Zhao, Darko Zikic, Marcel Prastawa, Mauricio Reyes, and Koenraad Van Leemput. The multimodal brain tumor image segmentation benchmark (brats). IEEE Transactions on Medical Imaging, 34:1993–2024, 2015.

[MKH19] Rafael Müller, Simon Kornblith, and Geoffrey E. Hinton. When does label smoothing help? ArXiv, abs/1906.02629, 2019.

[MMH11] Yuriy Mileyko, Sayan Mukherjee, and John Harer. Probability measures on the space of persistence diagrams. Inverse Problems, 27:124007, 2011.

[MMQ+16] Syed Osama Mahboob, Rachael McPhillips, Zhen Qiu, Yun Jiang, Carl Meggs, Giuseppe Schiavone, Tim W. Button, Marc P. Y. Desmulliez, Christine Demore, Sandy Cochran, and Sam Eljamel. Intraoperative ultrasound-guided resection of gliomas: A meta-analysis and review of the literature. World neurosurgery, 92:255–263, 2016.

[MPL10] Thomas E. Merchant, Ian F. Pollack, and Jay S. Loeffler. Brain tumors across the age spectrum: biology, therapy, and late effects. Seminars in radiation oncology, 20 1:58–66, 2010.

[MRMK20] Gaby N. Moawad, Sara Rahman, Martin A. Martino, and Jordan S. Klebanoff. Robotic surgery

during the covid pandemic: why now and why for the future. Journal of Robotic Surgery, 14:917 – 920, 2020.

[MRS+23] Matia Martucci, Rosellina Russo, Francesco Schimperna, Gabriella D'Apolito, Marco Panfili, Alessandro Grimaldi, Alessandro Perna, Andrea Maurizio Ferranti, Giuseppe Varcasia, Carolina Giordano, and Simona Gaudino. Magnetic resonance imaging of primary adult brain tumors: State of the art and future perspectives. Biomedicines, 11, 2023.

[MSM+20] Eleanor C. Mackle, Jonathan Shapey, Efthymios Maneas, Shakeel R. Saeed, Robert Bradford, Sebastien Ourselin, Tom Kamiel Magda Vercauteren, and Adrien E. Desjardins. Patient-specific polyvinyl alcohol phantom fabrication with ultrasound and x-ray contrast for brain tumor surgery planning. Journal of visualized experiments : JoVE, 2020.

[MSR+22] Alejandra Mosteiro, Alberto Di Somma, Pedro Roldán Ramos, Abel Ferrés, Andrea De Rosa, Sofía González-Ortiz, Joaquim Enseñat, and José Juan González. Is intraoperative ultrasound more efficient than magnetic resonance in neurosurgical oncology? an exploratory cost-effectiveness analysis. Frontiers in Oncology, 12, 2022.

[MT06] Oleg V. Michailovich and Allen Tannenbaum. Despeckling of medical ultrasound images. IEEE Transactions on Ultrasonics, Ferroelectrics and Frequency Control, 53:64–78, 2006.

[MTP07] Helen Morehouse, Harshada Pranav Thaker, and Chandowti Persaud. Addition of metamucil to gelatin for a realistic breast biopsy phantom. JUM, 26(8):1123–1126, 2007.

[Muk17] Siddhartha Mukherjee. Ai versus md: What happens when diagnosis is automated? The New Yorker, 3, 2017.

[Mur12] Kevin P. Murphy. Machine learning - a probabilistic perspective. In Adaptive computation and machine learning series, 2012.

[Nat21] National Institute for Health and Care Excellence. Brain tumours (primary) and brain metastases in over 16s: Recommendations, January 2021.

[Neu51] Von Neumann. Various techniques used in connection with random digits. Notes by GE Forsythe, pages 36–38, 1951.

[NHCD16] Laura Nolting, Patrick S Hunt, Thomas Cook, and Barton Douglas. An inexpensive and easy ultrasound phantom. Journal of Ultrasound in Medicine, 35, 2016.

[NHS25] NHS England. Millions to benefit from nhs robot drive. https://www.england.nhs.uk/2025/06/millions-to-benefit-from-nhs-robot-drive/, June 2025.

[NS56] Allen Newell and Herbert A. Simon. The logic theory machine-a complex information processing system. IRE Trans. Inf. Theory, 2:61–79, 1956.

[OKN+10] Shiro Ohue, Yoshiaki Kumon, Shigeyuki Nagato, Shohei Kohno, Hironobu Harada, Kou Nakagawa, Keiichi Kikuchi, Hitoshi Miki, and Takanori Ohnishi. Evaluation of intraoperative brain shift using an ultrasound-linked navigation system for brain tumor surgery. Neurologia medico-chirurgica, 50(4):291–300, 2010.

[Ope22] OpenAI. Introducing chatgpt, November 2022.

[Orc18] Guy H. Orcutt. A new regression analyser. Journal of the Royal Statistical Society, 111(1):54–70, 12 2018.

[PAB+14] Johan Pallud, Etienne Audureau, Marie Blonski, Nader Sanai, Luc Bauchet, Denys Fontaine, Emmanuel Mandonnet, Edouard Dezamis, Dimitri Psimaras, Jacques Guyotat, et al. Epileptic seizures in diffuse low-grade gliomas in adults. Brain, 137(2):449–462, 2014.

[PBEe22] Sarthak Pati, Ujjwal Baid, Brandon Edwards, and et.al. Federated learning enables big data for rare cancer boundary detection. Nature Communications, 13, 2022.

[PC23] Eva Pachetti and Sara Colantonio. A systematic review of few-shot learning in medical imaging. Artificial intelligence in medicine, 156:102949, 2023.

[PCC+22] Puneet Plaha, Sophie Camp, Jonathan Cook, Peter McCulloch, Natalie Voets, Ruichong Ma, Martin J B Taphoorn, Linda Dirven, Matthew Grech-Sollars, Colin Watts, Helen Bulbeck, Michael D Jenkinson, Matthew Williams, Adrian Lim, Luke Dixon, Stephen John Price, Keyoumars Ashkan, Vasileios Apostolopoulos, Vicki S Barber, Amy Taylor, FUTURE-GB collaborators, and Dipankar Nandi. Future-gb: functional and ultrasound-guided resection of glioblastoma – a two-stage randomised control trial. BMJ Open, 12(11), 2022.

[PD07] Florent Perronnin and Christopher R. Dance. Fisher kernels on visual vocabularies for image categorization. 2007 IEEE Conference on Computer Vision and Pattern Recognition, pages 1–8, 2007.

[PG11] Stephen J. Price and Jonathan H. Gillard. Imaging biomarkers of brain tumour margin and tumour invasion. The British journal of radiology, 84 Spec No 2:S159–67, 2011.

[PGM+19] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit

Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In Advances in Neural Information Processing Systems 32, pages 8024–8035. Curran Associates, Inc., 2019.

[PJM+20] Vitali Petsiuk, R. Jain, Varun Manjunatha, Vlad I. Morariu, Ashutosh Mehra, Vicente Ordonez, and Kate Saenko. Black-box explanation of object detectors via saliency maps. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 11438–11447, 2020.

[PL16] Allen Perkins and Gerald Liu. Primary brain tumors in adults: Diagnosis and treatment. American family physician, 93 3:211–7, 2016.

[PLLZ21] Bin Pu, Kenli Li, Shengli Li, and Ningbo Zhu. Automatic fetal ultrasound standard plane recognition based on deep learning and iiot. IEEE Transactions on Industrial Informatics, 17:7771–7780, 2021.

[PMM+15] Dino Podlesek, Tobias Meyer, Ute Morgenstern, Gabriele Schackert, and Matthias Kirsch. Improved visualization of intracranial vessels with intraoperative coregistration of rotational digital subtraction angiography and intraoperative 3d ultrasound. PLoS One, 10(3):e0121345, 2015.

[PMVS22] Dimitrios Psychogyios, Evangelos Mazomenos, Francisco Vasconcelos, and Danail Stoyanov. Msdesis: Multi-task stereo disparity estimation and surgical instrument segmentation. IEEE transactions on medical imaging, 41:3218 – 3230, 2022.

[PP24] Catalina Poggi and Martin Palavecino. Ultrasound principles and instrumentation. Surgery Open Science, 18:123 – 128, 2024.

[PPD+21] Rafael Padilla, Wesley L. Passos, Thadeu L. B. Dias, Sergio L. Netto, and Eduardo Antônio Barros da Silva. A comparative analysis of object detection metrics with a companion open-source toolkit. Electronics, 2021.

[PTD+22] Walter Hugo Lopez Pinaya, Petru-Daniel Tudosiu, Jessica Dafflon, Pedro F. Da Costa, Virginia Fernandez, Parashkev Nachev, Sébastien Ourselin, and Manuel Jorge Cardoso. Brain imaging generation with latent diffusion models. ArXiv, abs/2209.07162, 2022.

[QL20] Guo-Jun Qi and Jiebo Luo. Small data challenges in big data era: A survey of recent progress on unsupervised and semi-supervised methods. IEEE transactions on pattern analysis and machine intelligence, PP, 2020.

[QWK+18] Syed A. Quadri, Muhammad Waqas, Inamullah Khan, Muhammad Adnan Khan, Sajid S. Suriya, Mudassir Farooqui, and Brian Fiani. High-intensity focused ultrasound: past, present, and future in neurosurgery. Neurosurgical focus, 44 2:E16, 2018.

[RC15] Hassan Rivaz and D. Louis Collins. Deformable registration of preoperative mr, pre-resection ultrasound, and post-resection ultrasound images of neurosurgery. International Journal of Computer Assisted Radiology and Surgery, 10:1017–1028, 2015.

[RCEG24] Alfie Roddan, Tobias Czempiel, Daniel S. Elson, and Stamatia Giannarou. Calibration-jitter: Augmentation of hyperspectral data for improved surgical scene segmentation. Healthcare Technology Letters, 11:345 – 354, 2024.

[RCR24] Clinical radiology, workforce census 2024, 2024.

[RDGF15] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 779–788, 2015.

[RDS+14] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. International Journal of Computer Vision, 115:211 – 252, 2014.

[RFB15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In MICCAI, 2015.

[RFKL19] Aravind Rajeswaran, Chelsea Finn, Sham M. Kakade, and Sergey Levine. Meta-learning with implicit gradients. In Neural Information Processing Systems, 2019.

[RHGS15] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39:1137–1149, 2015.

[RHW86] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. Nature, 323:533–536, 1986.

[RKH+21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In International Conference on Machine Learning, 2021.

[RLR+07] I-A Rasmussen, F Lindseth, OM Rygh, EM Berntsen, T Selbekk, J Xu, TA Nagelhus Hernes, E Harg, A Håberg, and G Unsgaard. Functional neuronavigation combined with intraoperative 3d ultrasound: initial experiences during surgical resections close to eloquent brain areas and future directions in automatic brain shift compensation of preoperative data. Acta neurochirurgica, 149:365–378, 2007.

[RMJ+23] Alireza Rafiei, Ronald Moore, Sina Mohazzeb Jahromi, Farshid Hajati, and Rishikesan Kamaleswaran. Meta-learning in healthcare: A survey. SN Comput. Sci., 5:791, 2023.

[RMO+20] Subhankar Roy, Willi Menapace, Sebastiaan Oei, Ben Luijten, Enrico Fini, Cristiano Saltori, Iris A. M. Huijben, Nishith Chennakeshava, Federico Mento, Alessandro Sentelli, Emanuele Peschiera, Riccardo Trevisan, Giovanni Maschietto, Elena Torri, Riccardo Inchingolo, Andrea Smargiassi, Gino Soldati, Paolo Rota, Andrea Passerini, Ruud J. G. van Sloun, Elisa Ricci, and Libertario Demi. Deep learning for classification and localization of covid-19 markers in point-of-care lung ultrasound. IEEE Transactions on Medical Imaging, 39:2676–2687, 2020.

[RMS19] Prem K. Rachakonda, Bala Muralikrishnan, and Daniel Sawyer. Sources of errors in structured light 3d scanners. In Defense + Commercial Sensing, 2019.

[Ros57] Frank Rosenblatt. The perceptron: A perceiving and recognizing automaton. Report, Project PARA, Cornell Aeronautical Laboratory, January 1957.

[RPV+21] Francesco Restelli, Bianca Pollo, Ignazio Gaspare Vetrano, Samuele Cabras, Morgan Broggi, Marco Schiariti, Jacopo Falco, Camilla de Laurentis, Gabriella Raccuia, Paolo Ferroli, and Francesco Acerbi. Confocal laser microscopy in neurosurgery: State of the art of actual clinical applications. Journal of Clinical Medicine, 10, 2021.

[RPW18] Audrey Repetti, Marcelo Pereyra, and Yves Wiaux. Scalable bayesian uncertainty quantification in imaging inverse problems via convex optimization. SIAM J. Imaging Sci., 12:87–118, 2018.

[RS23] E Rough and N Sutherland. Review of brain tumour research funding, 2023.

[RTG+19] Seyed Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian D. Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 658–666, 2019.

[Rud16] Sebastian Ruder. An overview of gradient descent optimization algorithms. ArXiv, abs/1609.04747, 2016.

[RZKB19] Maithra Raghu, Chiyuan Zhang, Jon M. Kleinberg, and Samy Bengio. Transfusion: Understanding transfer learning for medical imaging. In Neural Information Processing Systems, 2019.

[Sam59] A. L. Samuel. Some studies in machine learning using the game of checkers. IBM Journal of Research and Development, 3(3):210–229, 1959.

[SBB+10] Tait D. Shanafelt, Charles M. Balch, Gerald J. Béchamps, Thomas R. Russell, Lotte N. Dyrbye, Daniel V. Satele, Paul E. Collicott, Paul J. Novotny, Jeff Sloan, and Julie Ann Freischlag. Burnout and medical errors among american surgeons. Annals of Surgery, 251:995–1000, 2010.

[SBD+10] Andreas Stadlbauer, Michael Buchfelder, Marc T. Doelken, Thilo Hammen, and Oliver Ganslandt. Magnetic resonance spectroscopic imaging for visualization of the infiltration zone of glioma. Central European Neurosurgery, 72:63 – 69, 2010.

[SBP+17] Rahul A. Sastry, Wenya Linda Bi, Steven D. Pieper, Sarah F. Frisken, Tina Kapur, William M. Wells, and Alexandra J. Golby. Applications of ultrasound in the resection of brain tumors. Journal of Neuroimaging, 27, 2017.

[ŠBT+22] Andrej Šteňo, Ján Buvala, David Toma, Michaela Jezberová, and Juraj Šteňo. Navigated 3d-ultrasound power doppler and visualization of lenticulostriate arteries during resections of insular gliomas. Brain & Spine, 2, 2022.

[SCD20] Kyle H. Sheetz, Jake Claflin, and Justin B Dimick. Trends in the adoption of robotic surgery for common surgical procedures. JAMA Network Open, 3, 2020.

[Sch21] J. Schell. Projectors. https://github.com/Ocupe/Projectors, 2021.

[SCJ+12] Felix Sahm, David Capper, Astrid Jeibmann, Antje Habel, Werner Paulus, Dirk Troost, and Andreas Von Deimling. Addressing diffuse glioma as a systemic brain disease with single-cell analysis. Archives of neurology, 69(4):523–526, 2012.

[SDV+16] Ramprasaath R. Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. International Journal of Computer Vision, 128:336 – 359, 2016.

[SEG17] Seyed Sadegh Mohseni Salehi, Deniz Erdoğmuş, and Ali Gholipour. Tversky loss function for image segmentation using 3d fully convolutional deep networks. In MLMI@MICCAI, 2017.

[SF11] Nathan Silberman and Rob Fergus. Indoor scene segmentation using a structured light sensor. 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), pages 601–608, 2011.

[SGRwHK23] Moslem Sadeghi-Goughari, Hossein Rajabzadeh, Jeong woo Han, and Hyock-Ju Kwon. Artificial intelligence-assisted ultrasound-guided focused ultrasound therapy: a feasibility study. International Journal of Hyperthermia, 40, 2023.

[SHK+14] Daniel Scharstein, Heiko Hirschmüller, York Kitajima, Greg Krathwohl, Nera Nesic, Xi Wang, and Porter Westling. High-resolution stereo datasets with subpixel-accurate ground truth. In GCPR, 2014.

[SHSM21] Filip Suligoj, Christoff Marthinus Heunis, Jakub Sikorski, and Sarthak Misra. Robust–an autonomous robotic ultrasound system for medical imaging. IEEE Access, 9:67456–67465, 2021.

[SIBK18] Pradipta Sasmal, Yuji Iwahori, Manas Kamal Bhuyan, and Kunio Kasugai. Active contour segmentation of polyps in capsule endoscopic images. 2018 International Conference on Signals and Systems (ICSigSys), pages 201–204, 2018.

[Sin] Graham Singer. History of the modern graphics processor, part 2. https://www.techspot.com/article/653-history-of-the-gpu-part-2/. Accessed: 2025-7-11.

[SK18] Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. Advances in neural information processing systems, 31, 2018.

[SKB+18] Ulrich Seibold, Bernhard Kübler, Thomas Bahls, Robert Haslinger, and Florian Steidle. The DLR MiroSurge surgical robotic demonstrator. volume 1 of The Encyclopedia of Medical Robotics, pages 111–142. WORLD SCIENTIFIC, October 2018.

[SMR+18] Harsimrat Singh, Hemel Narendra Modi, Samriddha Ranjan, James Dilley, Dimitrios Airantzis, Guang-Zhong Yang, Ara Darzi, and Daniel Richard Leff. Robotic surgery improves technical performance and enhances prefrontal activation during high temporal demand. Annals of Biomedical Engineering, 46:1621 – 1636, 2018.

[SNH+23] Raphael Schäfer, Till Nicke, Henning Höfener, Annkristin Lange, Dorit Merhof, Friedrich Feuerhake, Volkmar Schulz, Johannes Lotz, and Fabian Kiessling. Overcoming data scarcity in biomedical imaging with a foundational multi-task model. Nature Computational Science, 4:495 – 509, 2023.

[SNH+24] Krishnakant Singh, Thanush Navaratnam, Jannik Holmer, Simone Schaub-Meyer, and Stefan Roth. Is synthetic data all we need? benchmarking the robustness of models trained with synthetic images. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pages 2505–2515, 2024.

[SOB+96] Stephen S. Skirboll, George A. Ojemann, Mitchel S. Berger, Ettore Lettich, and H. Richard Winn. Functional cortex and subcortical white matter located within gliomas. Neurosurgery, 38 4:678–84; discussion 684–5, 1996.

[SOPH16] Fabio Alexandre Spanhol, Luiz Oliveira, Caroline Petitjean, and Laurent Heutte. Breast cancer histopathological image classification using convolutional neural networks. 2016 International Joint Conference on Neural Networks (IJCNN), pages 2560–2567, 2016.

[SPB04] Joaquim Salvi, Jordi Pagès, and Joan Batlle. Pattern codification strategies in structured light systems. Pattern Recognit., 37:827–849, 2004.

[SRDS17] Daniel C. Stewart, Andres Rubiano, Kyle A. Dyson, and Chelsey S. Simmons. Mechanical characterization of human brain tumors from patients and comparison to potential surgical phantoms. PLoS ONE, 12, 2017.

[SRG+16] Hoo-Chang Shin, Holger R. Roth, Mingchen Gao, Le Lu, Ziyue Xu, Isabella Nogues, Jianhua Yao, Daniel J. Mollura, and Ronald M. Summers. Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. Ieee Transactions on Medical Imaging, 35:1285 – 1298, 2016.

[SS03a] D. Scharstein and R. Szeliski. High-accuracy stereo depth maps using structured light. In 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings., volume 1, pages I–I, 2003.

[SS03b] Daniel Scharstein and Richard Szeliski. High-accuracy stereo depth maps using structured light. 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings., 1:I–I, 2003.

[SSG+24] Daniel Susser, Daniel S Schiff, Sara Gerke, Laura Y. Cabrera, I. Glenn Cohen, Megan Doerr, Jordan Harrod, Kristin Kostick-Quenet, Jasmine McNealy, Michelle N Meyer, W. Nicholson Price, and Jennifer K Wagner. Synthetic health data: Real ethical promise and peril. The Hastings Center Report, 54:8 – 13, 2024.

[SSSM20] Vikas Kumar Singh, Salman Ahmed Shaikh, Prakash Shetty, and Aliasgar V. Moiyadi. Customized low cost model for hands-on training in intraoperative ultrasound for neurosurgeons: Our experience and review of literature. World neurosurgery, 2020.

[SSV+20] Richard Shaw, Carole H. Sudre, Thomas Varsavsky, Sébastien Ourselin, and Manuel Jorge Cardoso. A k-space model of movement artefacts: Application to segmentation augmentation and artefact removal. IEEE Transactions on Medical Imaging, 39:2881–2892, 2020.

[ST16] Philipp Steiger and Harriet C. Thoeny. Prostate mri based on pi-rads version 2: how we review and report. Cancer Imaging, 16, 2016.

[SVI+15] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2818–2826, 2015.

[SVS+20] Sadaf Soloukey, Arnaud J. P. E. Vincent, Djaina D Satoer, Frits Mastik, Marion Smits, Clemens M F Dirven, Christos Strydis, Johannes G. Bosch, Antonius F. W. van der Steen, Chris I. De Zeeuw, Sebastiaan K. E. Koekkoek, and Pieter Kruizinga. Functional ultrasound (fus) during awake brain surgery: The clinical potential of intra-operative functional and vascular brain mapping. Frontiers in Neuroscience, 13, 2020.

[SWH+19] George Shih, Carol C. Wu, Safwan S. Halabi, Marc D. Kohli, Luciano M. Prevedello, Tessa S Cook, Arjun Sharma, Judith Amorosa, Veronica A. Arteaga, Maya Galperin-Aizenberg, Ritu R. Gill, Myrna C. B. Godoy, Stephen Hobbs, Jean Jeudy, Archana T. Laroia, Palmi N Shah, Dharshan R Vummidi, Kavitha Yaddanapudi, and Anouk Stein. Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. Radiology. Artificial intelligence, 1 1:e180041, 2019.

[SZ15] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In International Conference on Learning Representations, 2015.

[SZC+20] Trevor Scott Standley, Amir Roshan Zamir, Dawn Chen, Leonidas J. Guibas, Jitendra Malik, and Silvio Savarese. Which tasks should be learned together in multi-task learning? In ICML, 2020.

[Sze11] Richard Szeliski. Computer vision - algorithms and applications. In Texts in Computer Science, 2011.

[SZP+22] Benjamin Saß, Darko Zivkovic, Mirza Pojskic, Christopher Nimsky, and Miriam HA Bopp. Navigated intraoperative 3d ultrasound in glioblastoma surgery: analysis of imaging features and impact on extent of resection. Frontiers in Neuroscience, 16:883584, 2022.

[Tam21] Vethunan Tamalvanan. Foreseeable challenges in developing telesurgery for low income and middle-income countries. International Surgery Journal, 2021.

[TBT+19] Gianluca Trevisi, Paolo Barbone, Giorgio Treglia, Maria Vittoria Mattoli, and Annunziato Mangiola. Reliability of intraoperative ultrasound in detecting tumor residual after brain diffuse glioma surgery: a systematic review and meta-analysis. Neurosurgical Review, 43:1221 – 1233, 2019.

[TDM+23] Junjiao Tian, Xiaoliang Dai, Chih-Yao Ma, Zecheng He, Yen-Cheng Liu, and Zsolt Kira. Trainable projected gradient method for robust fine-tuning. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 7836–7845, 2023.

[TF14] Mickael Tanter and Mathias Fink. Ultrafast imaging in biomedical ultrasound. IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control, 61:102–119, 2014.

[TG22] Samyakh Tukra and Stamatia Giannarou. Stereo depth estimation via self-supervised contrastive representation learning. In International Conference on Medical Image Computing and Computer-Assisted Intervention, 2022.

[TLSK23] Junjiao Tian, Yen-Cheng Liu, James Seale Smith, and Zsolt Kira. Fast trainable projection for robust fine-tuning. ArXiv, abs/2310.19182, 2023.

[Tur50] Alan M. Turing. Computing machinery and intelligence. Mind, LIX:433–460, 1950.

[UL19] Geirmund Unsgård and Frank Lindseth. 3d ultrasound–guided resection of low-grade gliomas: principles and clinical examples. Neurosurgical focus, 47(6):E9, 2019.

[UOM+02] Geirmund Unsgaard, Steinar Ommedal, Tomm Muller, Aage Gronningsaeter, and Toril A Nagelhus Hernes. Neuronavigation by intraoperative three-dimensional ultrasound: initial experience during brain tumor resection. Neurosurgery, 50(4):804–812, 2002.

[UpS24] UpSurgeOn. Braintumorbox. https://store.upsurgeon.com/products/braintumorbox/, 2024.

[USHE18] Hristina Uzunova, Sandra Schultz, Heinz Handels, and Jan Ehrhardt. Unsupervised pathology detection in medical images using conditional variational autoencoders. International Journal of Computer Assisted Radiology and Surgery, 14:451–461, 2018.

[USS+17] Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. Sparsity invariant cnns. In International Conference on 3D Vision (3DV), 2017.

[VC64] V. N. Vapnik and A. Ya. Chervonenkis. A class of algorithms for pattern recognition learning. Avtomatika i Telemekhanika, 25(6):937–945, 1964.

[vdHdBdKG18] Thomas L. A. van den Heuvel, Dagmar de Bruijn, Chris L. de Korte, and Bram van Ginneken. Automated measurement of fetal head circumference using 2d ultrasound images. PLOS ONE, 13(8):1–20, 08 2018.

[VDWSN+14] Stéfan Van Der Walt, Johannes L. Schönberger, Juan Nunez-Iglesias, François Boulogne, Joshua D. Warner, Neil Yager, Emmanuelle Gouillart, Tony Yu, and the scikit-image contributors. scikit-image: image processing in Python. PeerJ, 2:e453, 6 2014.

[VGO+20] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods, 17:261–272, 2020.

[VHOT24] Chris Varghese, Ewen M. Harrison, Greg O'Grady, and Eric J. Topol. Artificial intelligence in surgery. Nature medicine, 2024.

[VLNZ21] Maria Victorova, Michael Ka-Shing Lee, David Navarro-Alarcon, and Yong Ping Zheng. Follow the curve: Robotic ultrasound navigation with learning based localization of spinous processes for scoliosis assessment. IEEE Access, PP:1–1, 2021.

[VLW+18] Bastiaan S. Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant cnns for digital pathology. In International Conference on Medical Image Computing and Computer-Assisted Intervention, 2018.

[VPMH21] Anuja Vats, Marius Pedersen, Ahmed Kedir Mohammed, and Øistein Hovde. Learning more for free - a multi task learning approach for improved pathology classification in capsule endoscopy. In International Conference on Medical Image Computing and Computer-Assisted Intervention, 2021.

[VSFAPMF12] Gonzalo Vegas-Sánchez-Ferrero, Santiago Aja-Fernández, Cesar Palencia, and Marcos Martín-Fernández. A generalized gamma mixture model for ultrasonic tissue characterization. Computational and Mathematical Methods in Medicine, 2012, 2012.

[VSP+17] Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Neural Information Processing Systems, 2017.

[VST+22] Bennett VanBerlo, Delaney Smith, Jared Tschirhart, Blake Vanberlo, Derek Wu, Alex Ford, Joseph McCauley, Benjamin Wu, Rushil Chaudhary, Chintan Dave, Jordan Ho, Jason L.

Deglint, Brian Li, and Robert T. Arntfield. Enhancing annotation efficiency with machine learning: Automated partitioning of a lung ultrasound dataset by view. Diagnostics, 12, 2022.

[VSV20] Nidhi Vyas, Shreya Saxena, and Thomas C. Voice. Learning soft labels via meta learning. ArXiv, abs/2009.09496, 2020.

[VZE+16] Salvatore Virga, Oliver Zettinig, Marco Esposito, Karin Pfister, Benjamin Frisch, Thomas Neff, Nassir Navab, and Christoph Hennersperger. Automatic force-compliant robotic ultrasound screening of abdominal aortic aneurysms. 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages 508–513, 2016.

[WCL+17] Lingyun Wu, Jie-Zhi Cheng, Shengli Li, Baiying Lei, Tianfu Wang, and Dong Ni. Fuiqa: Fetal ultrasound image quality assessment with deep convolutional networks. IEEE Transactions on Cybernetics, 47:1336–1349, 2017.

[WDA+23a] Alistair Weld, L. Dixon, G. Anichini, N. Patel, A. Nimer, M. Dyck, K. ONeill, A. Lim, Stamatia Giannarou, and S. Camp. Challenges with segmenting intraoperative ultrasound for brain tumours. Acta Neurochirurgica, 166, 2023.

[WDA+23b] Alistair Weld, Luke Dixon, Giulio Anichini, Michael Dyck, Alex Ranne, Sophie J. Camp, and Stamatia Giannarou. Identifying visible tissue in intraoperative ultrasound images during brain surgery: A method and application. ArXiv, abs/2306.01190, 2023.

[Wer82] Paul J. Werbos. Applications of advances in nonlinear sensitivity analysis. In Lecture Notes in Control and Information Sciences. System Modeling and Optimization, pages 762–770. Springer Berlin Heidelberg, 1982.

[WKH+20] Martin J. Willemink, Wojciech A Koszek, Cailin Hardell, Jie Wu, Dominik Fleischmann, Hugh Harvey, Les R. Folio, Ronald M. Summers, D. Rubin, and Matthew P. Lungren. Preparing medical imaging data for machine learning. Radiology, page 192224, 2020.

[WLF+13] Wolfgang Wein, Alexander Ladikos, Bernhard Fuerst, Amit Shah, Kanishka Sharma, and Nassir Navab. Global registration of ultrasound to mri using the lc2 metric for enabling neurosurgical guidance. Medical image computing and computer-assisted intervention : MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention, 16 Pt 1:34–41, 2013.

[WNBL17] Juliana Wilson, Lorraine Ng, Vaughn A. Browne, and Resa E Lewiss. An easy-to-make, low-cost ultrasound phantom for simulation training in abscess identification and aspiration. Journal of Ultrasound in Medicine, 36, 2017.

[WRR+22] Urs Würtemberger, Alexander Rau, Marco Reisert, Elias Kellner, Martin Diebold, Daniel Erny, Peter Christoph Reinacher, Jonas Aurel Hosp, Marc Hohenhaus, Horst Urbach, and Theo Demerath. Differentiation of perilesional edema in glioblastomas and brain metastases: Comparison of diffusion tensor imaging, neurite orientation dispersion and density imaging and diffusion microstructure imaging. Cancers, 15, 2022.

[WSB03] Zhou Wang, Eero P. Simoncelli, and Alan Conrad Bovik. Multiscale structural similarity for image quality assessment. The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003, 2:1398–1402 Vol.2, 2003.

[WTN+23] Andrew Awuah Wireko, Pearl O Tenkorang, Jyi Cheng Ng, Lian David, Rohan Yarlagadda, Toufik Abdul-Rahman, Vladyslav Sikora, and Arda Işık. Addressing africa's outrageous neurosurgeons deficit: what could the problem be? International Journal of Surgery (London, England), 109:1808 – 1809, 2023.

[Wu83] C. F. Jeff Wu. On the convergence properties of the em algorithm. Annals of Statistics, 11:95–103, 1983.

[WWH+24] Daniel Won, James Walker, Russ Horowitz, Sandeep Bharadwaj, Edward Carlton, and Helena Gabriel. Sound the alarm: The sonographer shortage is echoing across healthcare. Journal of ultrasound in medicine : official journal of the American Institute of Ultrasound in Medicine, 2024.

[XFU+17] Yiming Xiao, Maryse Fortin, Geirmund Unsgård, Hassan Rivaz, and Ingerid Reinertsen. Retrospective evaluation of cerebral tumors (resect): A clinical database of pre-operative mri and intra-operative ultrasound in low-grade glioma surgeries. Medical physics, 44 7:3875–3882, 2017.

[XHE22] Chi Xu, Baoru Huang, and Daniel S Elson. Self-supervised monocular depth estimation with 3-d displacement module for laparoscopic images. IEEE Transactions on Medical Robotics and Bionics, 4(2):331–334, 2022.

[XMW+19] Yiming Xiao, Andreas K. Maier, Wolfgang Wein, Roozbeh Shams, Samuel Kadoury, David Drobny, Marc Modat, Ingerid Reinertsen, Hassan Rivaz, Matthieu Chabanas, Maryse Fortin, Inês Machado, Yangming Ou, Mattias P. Heinrich, Julia Anne Schnabel, and Xia Zhong. Evaluation of mri to ultrasound registration methods for brain shift correction: The curious2018 challenge. IEEE Transactions on Medical Imaging, 39:777–786, 2019.

[Yin19] Xue Ying. An overview of overfitting and its solutions. Journal of Physics: Conference Series, 2019.

[YZL+21] Jiawei Yang, Yao Zhang, Yuan Liang, Yang Zhang, Lei He, and Zhiqiang He. Tumorcp: A simple but effective object-level data augmentation for tumor segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention, 2021.

[ZCG20] Jian Zhan, João Cartucho, and Stamatia Giannarou. Autonomous tissue scanning under free-form motion for intraoperative tissue characterisation. 2020 IEEE International Conference on Robotics and Automation (ICRA), pages 11147–11154, 2020.

[ZGCS23] Xiaohui Zhang, Ahana Gangopadhyay, Hsi-Ming Chang, and Ravi Soni. Diffusion model-based data augmentation for lung ultrasound classification with limited data. In ML4H@NeurIPS, 2023.

[ZLG98] Xuli Zong, Andrew F. Laine, and Edward A. Geiser. Speckle reduction and contrast enhancement of echocardiograms via multiscale nonlinear processing. IEEE Transactions on Medical Imaging, 17:532–540, 1998.

[ZqYL+23] Yi Zhang, Hong quan Yu, Yunqian Li, Haiyang Xu, Liu Yang, Peilin Shan, Yue-Wu Du, Xiaokai Yan, and Xuan Chen. Raman spectroscopy: A prospective intraoperative visualization technique for gliomas. Frontiers in Oncology, 12, 2023.

[ZSL+20] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. ArXiv, abs/2010.04159, 2020.

[ZSX+23] Xiandi Zhang, Zhifeng Shi, Yuanxin Xie, Yong Wang, Chao Shen, Zengxin Qi, Liqiong Zhang, Bojie Yang, Jinhua Yu, and Hong Ding. Quantitative analysis using intraoperative contrast-enhanced ultrasound in adult-type diffuse gliomas with isocitrate dehydrogenase mutations: association between hemodynamics and molecular features. Ultrasonography, 42(4):561, 2023.

[ZSZ+20] Chaoqiang Zhao, Qiyu Sun, Chongzhen Zhang, Yang Tang, and Feng Qian. Monocular depth estimation based on deep learning: An overview. Science China Technological Sciences, pages 1–16, 2020.

[ZTAG24] Yihang Zhou, Rebecca Towning, Zaid Awad, and Stamatia Giannarou. Image synthesis with class-aware semantic diffusion models for surgical scene segmentation. Healthcare Technology Letters, 12, 2024.

[ZWC+22] Yan Zhao, Xiuying Wang, Tontong Che, Guoqing Bao, and Shuyu Li. Multi-task deep learning for medical image computing and analysis: A review. Computers in biology and medicine, 153:106496, 2022.

[ZXU+23] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Jaspreet Kaur Bagga, Robert Tinn, Sam Preston, Rajesh N. Rao, Mu-Hsin Wei, Naveen Valluri, Cliff Wong, Matthew P. Lungren, Tristan Naumann, and Hoifung Poon. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. 2023.

Appendix A

A Background on Machine Learning Algorithms

A common theme that runs through this thesis and is also a significant factor in a large portion of the work in the medical/healthcare (and in our case surgical) domain is the limitation of data. This appendix chapter provides a short introduction to machine learning algorithms, keeping the reading light, with a focus on the history, starting from the late 19th century.

I hope that this section will provide context as to why I chose to explore some of the avenues of research I did and what parallel research could be explored by others. It is unquestionable that machine learning algorithms have become one of, if not the most important and defining technical subjects of this period; transformative of modern society. Nevertheless, there is clearly a disparity in advancements in areas of data abundance and data scarcity; where my PhD and similar work are firmly in the latter category. Not enough has been done within this community to make this transparent. Progress is often touted in the medical imaging domain; however, these examples are largely limited to very specific clinical tasks where there is data, such as preoperative MRI. But it is only a tiny subset of the overall set of clinical tasks. This is reflected in the limited clinical translation of machine learning algorithms into regular clinical environments.

A.1 Early Developments

Modern systems and computers were born in the early 20th century. Perhaps the strongest attributions for these developments should be given to the groups who made reality from the theory; Bell Labs and the global military efforts during World War 2 and the Cold War. The key fundamental breakthroughs that laid the foundations for this - limiting to only from the late 19th century for brevity - can be found in the works from scientists of the fields of information theory such as Claude Shannon and Alan Turing. In addition to the works in mathematical logic by Kurt Gödel, Bertrand Russell, and David Hilbert, to name a few. Many other topics could and perhaps should also be listed. In particular, within the fields of physics, material science, and electronics. Without these achievements, relevant algorithmic research would have no platform on which to advance and no way to show application. However, an introduction to these topics is far beyond the scope of this thesis; instead, I recommend the reading of books such as "A history of modern computing" by Paul Ceruzzi [Cer98].

Perhaps the first instances of machine learning algorithms that were implemented and run on hardware would be examples from the 1930s to 1940s in statistical inference, modelling, and linear regression. Most notably, following earlier works of Sir Francis Galton, Sir Ronald Fisher, and Karl Pearson. In the most uncomplimentary terms, these implementations could be viewed (with a modern lens) as no more than automated line fitting algorithms or simple probabilistic and decision-making algorithms. For example, using punched cards to solve linear differential equations [FS41], and to solve least squares estimates for linear equations [Orc18]. Where the true technical challenge would be the data recording and digitising, and leveraging the hardware available. Regardless of this, these algorithms and early breakthroughs are the origin of what is now one of the most defining technical developments of the 21st century.

During the mid century, three significant public developments were made that bridged theoretical artificial intelligence (AI) with practical implementation. I choose to leave out the

philosophical developments in AI during this period, which in themselves were critical in defining how we believe modern AI should process information, learn, and behave. For a reasonable starting point for philosophical AI, I point the reader towards "Computing Machinery and Intelligence" by Alan Turing [Tur50]. The first of these breakthroughs, while not being a machine learning algorithm itself, was the 1956 automated reasoning implementation of the "Logic Theorist" by Allen Newell, Cliff Shaw, and Herbert Simon [NS56]. A computer programme capable of automatically proving theorems through the use of symbolic logical expression and processing. Frank Rosenblatt's Perceptron [Ros57] was implemented and simulated on an IBM 704 in 1957/8, marking the first realisation of a trainable neural network and a cornerstone in deep learning history. Then in 1959, Arthur Samuel [Sam59], developed the first self-learning AI for games, playing checkers using a minimax strategy and a pruning method.

As computers became a more common feature in research laboratories, many of the learning methods that are still common tools today were first implemented on hardware. Although the fundamental origins of these algorithms often trace back much earlier, I focus here on the first documented implementation cases, which also helps to reduce the impact of subjectivity. Below is a brief, "rattled-off" list of key machine learning breakthroughs; I include this for historical context, which in my view is essential for understanding today's technical space.

K-Means was first implemented in 1957 at Bell Labs by Stuart Lloyd [Llo82]. However, this was only published in 1982 and what mainstreamed the popularity of K-means came through publications in the 1960s. Nevertheless, Lloyd's experiments can possibly be regarded as the first modern machine learning implementation that is still used today. Similarly, the first publication that outlining K-Nearest-Neighbour (KNN) was published as early as 1951 [FH89]. However, it was not until 1967 [CH67] that KNN was fully formalised and used more popularly. Decision trees are a broad family of algorithms that encompass many different designs, with conceptual roots that trace back to rule-based decision tables of ancient civilisations. The first computerised Decision Tree programme is likely to be attributed to William Belson, who in 1959 used punched cards to match population samples [Bel59]. The theoretical foundations of the Support Vector Machine (SVM) were developed by Vladimir Vapnik and Alexey Chervonenkis in a series of papers in the 1960s based around "generalised portraits", algorithms that maximise

the distance between data from two classes, an idea later recognised as the maximum-margin criterion of the SVM [VC64]. Whether or not this was already being implemented on hardware in Russia is unclear. However, the most significant follow-up development to the SVM was the introduction of the kernel trick to simplify the handling of non-linear boundary learning [BGV92], improving the scalability and feasibility of the implementation of SVMs. Learning from sequential data has strong roots in Markov chains. Methods based on this framework drew earlier interest, but it was not until Leonard Baum and his colleagues [BP66] formalised the Hidden Markov Model in a series of papers starting in 1966, that it was possible to use Markov chains to infer hidden/latent states and their sequence transitions. A seminal probabilistic machine learning algorithm, which with minimal modification could be made fully Bayesian, was the Expectation-Maximisation algorithm introduced in 1977 by Arthur Dempster, Nan Laird, and Donald Rubin [DLR18]. This represented a large conceptual shift from previous machine learning methods, as this algorithm provided an iterative framework of monotonic log likelihood convergence (as proved by CFJ Wu [Wu83]) to find maximum likelihood estimates for models involving latent variables and an underlying generative process. By modelling parameters with prior distributions, this can be treated as optimising the maximum a posteriori objective, providing the bridge to Bayesian learning.

A.2 Deep Learning

A subset of machine learning, which has come to define the current research period, is, of course, deep learning; a category of methods that involve the development of models containing structures of multiple layers of hidden parameters. The technical advances of deep learning have closely followed developments in the availability and computing power of graphical processing units (GPUs). Again, to reduce subjectivity and for the sake of brevity, we will introduce this topic by describing what I perceive as the key historical milestones; while noting that I will have ultimately left out developments which others would deem important. For example, I will not describe the earlier work in areas such as control theory, which laid strong foundations for updating parameters using gradients.

The key algorithmic development that allowed the optimisation of a large number of parameters in multiple layers is backpropagation. It was first described by Seppo Linnainmaa in his 1970 master's thesis "The representation of the cumulative rounding error of an algorithm as a Taylor expansion of the local rounding errors" [Lin70], where he introduced reverse-mode automatic differentiation using the chain rule to compute derivatives. The first application of this technique to a neural network-specifically a multilayer perceptron (MLP) - extending Rosenblatt's perceptron - was by Paul Werbos in 1982 [Wer82]. Most significantly for its eventual popularity, David Rumelhart, Geoffrey Hinton, and Ronald Williams formalised and demonstrated backpropagation as a tool for pattern recognition in their 1986 paper [RHW86].

In parallel to the developments in optimisation, research on the design of neural network architecture was equally critical to discovering models capable of exploiting large datasets. An early milestone was Hopfield's 1982 network [Hop82], which introduced an energy-based associative memory model. This has continued to provide significant inspiration in the design of neural networks that contain "memory", allowing the propagation of information across sequences or time. However, in the 1989 paper by Yann LeCun, it was demonstrated for the

first time that convolutional networks could be trained end-to-end using "large data sets" with backpropagation [LBD+89]. The Long Short-Term Memory (LSTM) was proposed in 1997 by Sepp Hochreiter and Jürgen Schmidhuber [HS97], which provided an architectural solution to training models containing memory, solving the problem of vanishing gradients in recurrent neural networks. Enabling models to learn long-term dependencies and improving performance on sequential tasks.

A period of time elapsed until the next milestone occurred. AlexNet [KSH12] was developed by Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton and submitted in 2012 at the ImageNet Large Scale Visual Recognition Challenge [RDS+14]. One of the first modern deep learning neural network applied for image classification, which won the challenge by $\approx 10\%$, a significant leap in classification performance compared to the other submissions that were based on hand-crafted features (the common approach at the time); such as the scale-invariant feature transform (SIFT) [Low99] and Fisher Vectors [PD07]. From this point on, the benefit of deep learning research became clear to the wider community. Without diminishing the achievements of researchers in algorithmic development, it must be stated that the impact of hardware power and availability was the largest contributing factor to this breakthrough. Notable are the two contributions by Nvidia: the creation of the (marketed) world's first GPU in 1999, the GeForce 256 [Sin], designed primarily for 3D video gaming and acceleration - it is worth noting that this was manufactured by Taiwan Semiconductor Manufacturing Company Limited, who themselves played an enormous role in hardware developments and availability; and their invention of Compute Unified Device Architecture (CUDA) [Che10] in 2007 which leveraged the parallel computation of GPUs. AlexNet was trained using two 3GB Nvidia GTX 580 GPUs and written in CUDA. In addition, the efforts made into the creation of large annotated public data sets, containing hundreds of thousands to millions of images, facilitated the development of algorithms for large-scale / large-distribution applications. The most poignant example, clearly, is the ImageNet [DDS+09] dataset itself. A slew of deep learning advancements have since followed. Notable neural network architectures would include: Generative Adversarial Networks (GAN) 2014 [GPAM+14], Residual Neural Network (ResNet) 2015 [HZRS15], Transformer 2017 [VSP+17], Bidirectional encoder representations from transformers (BERT) 2018

[DCLT19], Vision Transformer (ViT) 2020 [DBK+20], Diffusion Models 2020 [HJA20b].

This extraordinary acceleration in research and development has made the algorithmic application for medical and surgical tasks more likely to succeed. When applied successfully, deep learning methods have now unequivocally proven to significantly outperform alternatives that rely on priors or hand-crafted features. However, this success requires the availability of large training datasets - which is a persistent issue in the healthcare/medical domain.

Appendix B

Standardisation of Convex Ultrasound Data Through Geometric Analysis and Augmentation

B.1 Introduction

The role of ultrasound (US) in healthcare has increased greatly over the last three decades and is now the main imaging modality in the evaluation of many different organ systems and pathologies. The ability of US to provide dynamic imaging in real time, combined with its relative affordability and adaptability to various clinical settings, has solidified its position as a useful tool in US-guided procedures and surgical guidance [DLGS+22] [KHC+13] [LGK+21]. As there is a global increase in demand for medical imaging (in all modalities) coupled with a global shortage of experts [WWH+24], the research on technical solutions has also increased considerably, notably within the space of data-driven algorithms [JKA+22] [RMO+20] [ACB+19], robotics and control [JSN23].

Unlike other imaging modalities such as magnetic resonance imaging (MRI) [MJB+15], public US data is relatively disorganised and sparse; lacking standardised and benchmarking datasets.

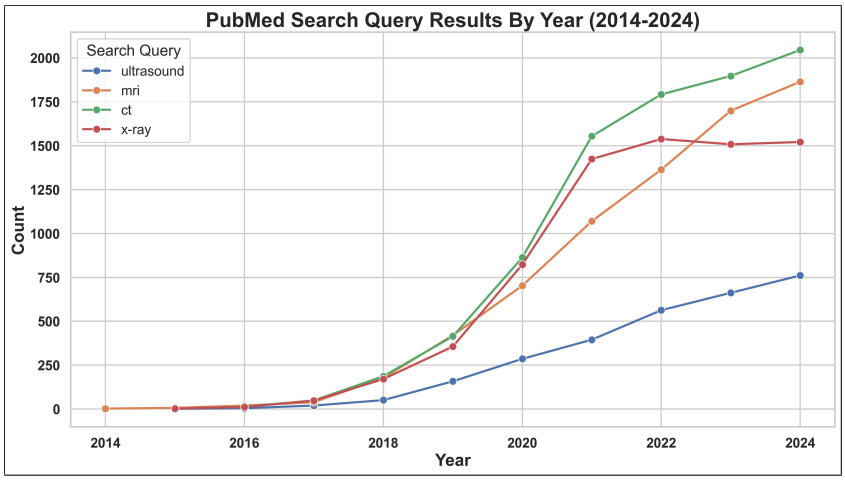


Figure B.1: Medical imaging modality deep learning research output trends.

In particular, for data-driven methods, this has presented a bottleneck in relevant research. This is a significant factor that has resulted in fewer US research output, relative to other modalities. For example, a simple search of `https://pubmed.ncbi.nlm.nih.gov` using the keyword queries "modality" AND "deep learning" shows that US lags far behind other modalities, as is shown in Fig. B.1. US images are highly variable depending on the machine used, the wide range of different probes, machine settings, capture methods, and operator dependence. The slight alterations in these variables may account for substantial differences in the US plane and image intensities. Since the publication of US data is often targeted at clinical audiences (e.g. `https://radiopaedia.org`), or consists of data collected by a single centre, there is no common consensus on standardising how to publish US data. This often leads to publishing images with manual modifications, such as cropping or the addition of additional text or graphical features. Particularly for convex US images, the variable curved geometry and decoupling of image height and scan line depth create additional data complexity and variability. These variation factors compound the difficulty of using public data for research and development.

From a technical perspective, this lack of standardisation can be perceived as equivalent to a lack of data. This has consequences for data-driven methods that either train on large volumes of data or that need to be evaluated on diverse datasets to prove robustness and generalisability. The option of manually collecting and annotating data is labour intensive and involves logistical variables such as ethics and data sharing between multiple centres. For example, the well-known breast US dataset BUSI [ADGKF19], which only contains 780 images, took about a year to

collect and annotate. Because of these reasons, synthetic phantoms are often substituted for in vivo data, at the cost of data realism. Or, research is published using private datasets exclusively, making reproducibility an issue. Technical methods for alleviating the issue of data scarcity and lack of standardisation are limited. The main solution proposed is the application of generative models for the generation of synthetic medical datasets [PTD+22] or for embedding into training frameworks where real images or annotations are limited or perhaps not available. However, there are many technical and ethical concerns and complications when it comes to the use of synthetic data in clinical applications [SSG+24] [GS23] [SNH+24].

In this work, an approach to standardise US datasets is proposed based on automatically extracting the US plane from convex US images. Simple use cases include normalising the position of the US plane within a 2D array to more advanced applications such as data augmentation for neural network pipelines [CZZ+24] or signal processing methods that process along scan lines [KWKN12b]. From a linear algebra perspective, performing matrix operations, such as convolutions or matrix multiplications, becomes more intuitive and logical when scan lines align with image columns. Therefore, having the convex plane extracted enables adjustment of these operations or the linearisation of the plane. A key limitation of these approaches is their reliance on the convex US plane being known a priori, where in the above two references manual labelling of the plane was done.

To our knowledge, no other work has focused on the specific task of automatically extracting the convex US plane. Methods have been explored to minimise manual annotation and data organisation labour. For example, determining the orientation and location of US scans relative to anatomical structures has been proposed to support diagnosis. These works often focus on the identification of the foetal standard plane [PLLZ21] [BKM+16]. Other works include partitioning of data sets based on the view of the lung [VST+22]. Additional topics include the automation of US quality assessment, which helps to ensure correct biological measurements [WCL+17], and could be also used for data set organisation. However, the limitations of these methods are the lack of explicit isolation of the relevant, full US-plane geometry and the limited use-case to specific anatomy or US-task. Furthermore, most proposed methods are based on deep learning, which introduces classic black-box interpretability and reliability concerns.

In this paper, we introduce a nonanatomy- or application-specific method for automatic extraction of the visible, convex US plane. This is done using traditional image processing techniques and representing the convex plane using annulus sector geometry. Our key contributions are as follows:

1) We propose the first framework for automatically extracting convex US planes from non-standardised US images, including those affected by common GUI elements, additional text and graphical features, and plane modifications such as cropping.

2) We demonstrate that convex US planes can be effectively represented using annulus sector geometry, enabling transformations that preserve the geometry of underlying features.

The first evaluation conducted is the accuracy and robustness of the proposed method in estimating the annulus sector parameters. This included an evaluation using both public and private datasets, where the ground-truth annotations were created manually. The data used for the evaluation included cases where there was incomplete visualisation of the convex plane and corruption. Further evaluation explored the viability of using the method to augment the image, including the invertibility of the process and measuring the deformation of a circular object post-augmentation.

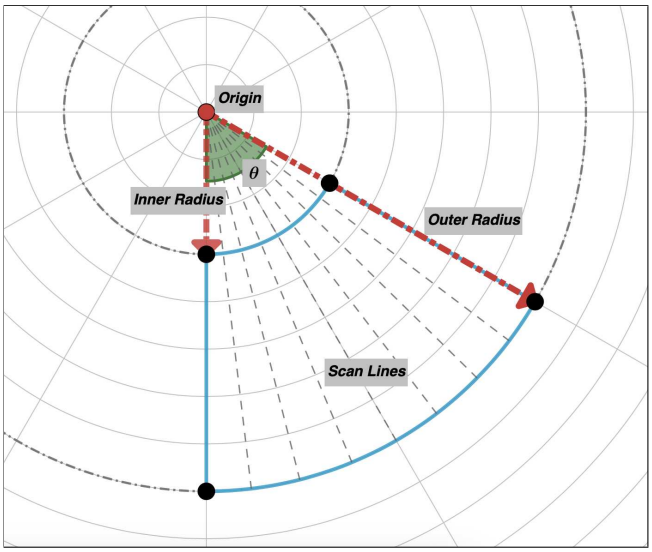


Figure B.2: Annulus sector geometric properties, created using `https://www.geogebra.org`.

B.2 Methodology

In this section, a description is given of the method for extracting the annulus sector parameters, highlighted in Fig. B.2 from convex US images, e.g. left subfigure in Fig. B.3. The proposed method can be broken down into four stages: 1) "Ultrasound Plane Masking", which aims to find the rough US plane and represent it as a binary mask, and to also remove the background GUI; 2) "Centre Of Plane Calculation" involves determining the centre line through the US plane; 3) "Radial Boundary Detection" estimating the radial boundaries of the plane, utilising the symmetry from the centre line; 4) "Annulus Sector Parameter Calculation", is the final stage of the algorithm where the annulus sector parameters are determined using the information from the previous stages. At the end of this section, the methodology for how the annulus sector parameters can be used for the automatic linearisation of a convex US plane is described. When required, we simplify the mathematical description with algorithmic notation, e.g. pythonic indexing or for loops such as [0] and [-1] for indexing the first and last element in a list, respectively.

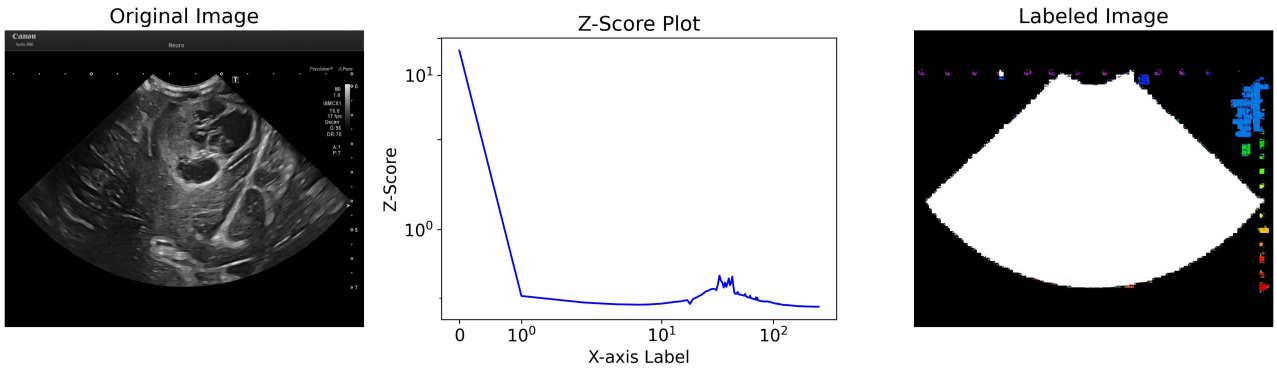


Figure B.3: Extracting the US plane. Left shows the original image. Middle shows a plot of the z-score, where in this circumstance there is only one spike as the background pixel is constant at 0. Right shows the output of the connected components process, where the US plane is the largest object and represented in white (the other colours represent other objects, belonging to the GUI).

B.2.1 Ultrasound Plane Masking

A raw 2D B-mode convex US image is defined as $I_{rgb} \in \mathbb{R}^{h,w,c}$ - an example of a standard US image with full GUI is shown in Fig. B.3. This image is converted into grayscale I_g. We define the image origin at the top left corner. The superimposed GUI on the image, for example, the header and text relating to machine and parameter settings, should be removed because they may affect further image processing. To identify the US plane within the image, a simple two-pass 4-connectivity connected component labelling method [HRG+17], [BB82] is used. First, the image is binarised to separate the foreground and black background pixels. Typically, the pixel intensity in the background, the area outside of the US plane, is equal to 0. However, this is not always the case due to variability on the data capture and storage. So, the background is detected by analysing the image's pixel intensity distribution.

$$H = hist(I_g, bins = [0, 255]) \tag{B.1}$$

As the background is homogeneous, its corresponding pixel intensities create a large spike within the intensity distribution. A spike in the histogram is determined if the frequency of an intensity is greater than an adaptive minimum threshold and is at least half the value of the highest bin frequency. This threshold ensures that only bins with sufficiently high frequencies

relative to the peak frequency are considered. The z-score analysis is used for spike detection. Since background pixels are homogeneous, they produce a distinct spike, defined sp_{thr}, in the histogram that meets our frequency threshold, regardless of high-intensity white pixels. We define the background intensity range using the furthest spike from the origin that satisfies the given criteria. Where the range is important to accommodate noise in the data capture or the saving. Determining the upper intensity of the background bg_{max} enables the creation of the foreground-background mask as:

$$Z[i] = \frac{H[i] - \mu}{\sigma}, \forall i \in [0..255] \tag{B.2}$$

$$\begin{aligned} sp_{thr} &= \max\left(\frac{\max(Z)}{2}, 2\right), \\ bg_{\max} &= \max\{\, i \in \{0, 1, 2, \ldots, 255\} : Z_i > sp_{thr}\}. \end{aligned} \tag{B.3}$$

$$I_b = \begin{cases} 1 & \text{if } I_g > bg_{max} \\ 0, & \text{otherwise} \end{cases} \tag{B.4}$$

where, μ and σ are the mean and standard deviation of the image's intensity distribution. Once the foreground-background mask I_b has been constructed, a connected component function can identify distinct objects by grouping adjacent pixels, separating the US plane from other foreground components in the image (e.g text, GUI components):

$$P[j, i] = \begin{cases} C_c & \text{if } f_{ccl}(I_b[j, i]) = C_c \\ 0, & \text{otherwise} \end{cases} \tag{B.5}$$

where, $f_{ccl}()$ denotes the two-pass, 4-connectivity connected component mapping. C_c denotes the connected components index where $C = \{C_1, C_2, \ldots, C_n\}$, and P is the output map of all connected components, that has the same dimensions as I_b. $f_{ccl}()$ works by assigning a pixel to a connected component C_n, if that pixel is contiguous either vertically or horizontally, to a pixel already belonging to C_n.

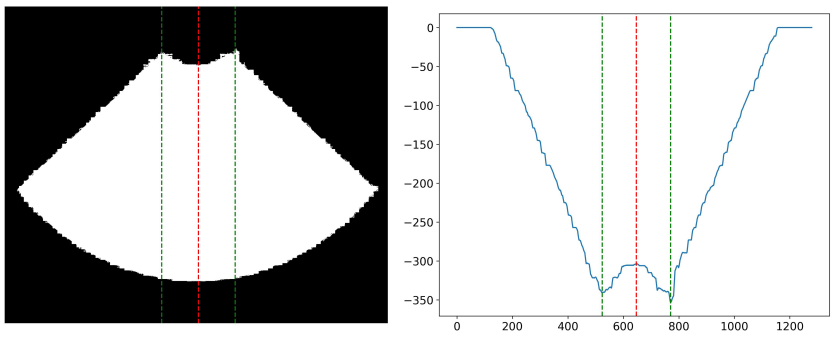


Figure B.4: (Right) An example of the calculation of the vertical axis of symmetry using the sliding window method. The blue line represents the intensities of vector S, the red line is the vertical line passing through point m and the green lines are the vertical lines through points min_l, min_r. (Left) The binarised US plane with the same lines superimposed.

The convex US image plane can then be isolated through element-wise multiplication with the largest connected component subset as:

$$I_c = I_b \odot \text{argmax}_i(|C_i|), i \in \{1, \ldots, n\} \tag{B.6}$$

where, I_c denotes the image containing only the US plane. An overview of the extraction of the US plane is shown with an example in Fig. B.3.

B.2.2 Centre Of Plane Calculation

The convex US image plane detected above can be represented as an annulus sector. The aim of this step is to use certain geometric assumptions to determine the parameters of this annulus sector such as the vertical axis of symmetry as shown in Fig. B.4. The key to ensuring robustness of the representation of the US plane is to determine the true edges of the annulus sector, which may be affected by noise and artefacts. Assuming increasing signal attenuation with penetration depth and near-field echogenicity, the top corners of the annulus sector and the centre point of the inner arc, shown in Fig. B.2, are calculated first. To determine these 3 key points, a horizontal window $\hat{S}$ (with a height of one pixel and a width equal to that of the image) is incrementally slid over the image, moving only along the vertical direction, from the

top of the image, to detect the inner arc.

$$\hat{S}^0 = [0]_{1\times w}$$
$$\hat{S}^j = \hat{S}^{j-1} - I_c[j,:], j = [1,\ldots,h/2], \hat{S} \leq 0 \tag{B.7}$$

$$S = \hat{S}^{h/2} * [0.2, 0.2, 0.2, 0.2, 0.2] \tag{B.8}$$

where, w and h are the width and height of I_c, respectively and $*$ denotes convolution. The range of half the height has been arbitrarily chosen. The vector S represents the accumulated row difference and is shown with the blue line in the right subfigure of Fig. B.4.

The centre of mass of the US plane is calculated to find an approximate of the centre of the inner curve. This is an intermediate step to prevent a negative impact in case there is asymmetry in the masked image I_c, which is likely due to image noise or superimposed GUI. An initial estimate of the x-coordinate of the centre of mass, $\hat{m}$, is calculated using the weighted average trapezoid rule as:

$$\hat{m} = \frac{\sum_{i=1}^{w-1} \frac{iS[i]+(i+1)S[i+1]}{2}}{\sum_{i=1}^{w-1} \frac{S[i]+S[i+1]}{2}} \tag{B.9}$$

Once $\hat{m}$ has been determined, the central convex hull is identified by locating the minima on either side (horizontal) of the centre of mass (displayed using the green vertical bars in Fig. B.4).

$$min_l = \min(S[i] : i = 1..\hat{m}),$$
$$min_r = \min(S[i] : i = \hat{m}+1..w) \tag{B.10}$$

To account for noise that may affect the mask, which would impact the convex hull, the final centre-of-mass estimation is calculated by averaging the centre of mass with the midpoint of min_l and min_r.

$$m = \frac{\hat{m} + (0.5(min_r - min_l) + min_l)}{2} \tag{B.11}$$

Assuming that the US plane is normally orientated (without rotation), which is standard for 2D US, the coordinates of the inner and outer arc can be calculated by finding the first and final point of intersection of the US plane in I_c, along the vertical line passing through point m as:

$$\begin{aligned} S(m) &= \{\, j \in \{0, 1, \ldots, h-1\} \mid I_c(j, m) = 1 \,\}, \\ \text{arc}_{\text{inner}} &= \big(\min S(m),\, m\big), \quad \text{arc}_{\text{outer}} = \big(\max S(m),\, m\big). \end{aligned} \tag{B.12}$$

B.2.3 Radial Boundary Detection

The radial boundaries, i.e the legs of the convex plane of I_c, can be initially estimated using the farthest points, to the left and right of the x-coordinate of the centre of mass m:

$$\begin{aligned} \hat{\text{edges}}_l = \Bigg\{ \Big(j,\ \min\Big\{ i \in \{0, \ldots, m-1\} : I_c(j,i) = 1 \Big\}\Big) : \\ j \in \{0, \ldots, h-1\} \Bigg\}, \\ \hat{\text{edges}}_r = \Bigg\{ \Big(j,\ \max\Big\{ i \in \{0, \ldots, m-1\} : I_c(j,i) = 1 \Big\}\Big) : \\ j \in \{0, \ldots, h-1\} \Bigg\}. \end{aligned} \tag{B.13}$$

The angular vertices that connect the radial boundaries and the outer arc are determined to find the end coordinates of the radial boundary. By determining the vertices, the edge points associated with the radial boundary are isolated and can be used to represent a prior estimate of the radial boundaries, although affected by noise and discontinuities along the US plane.

$$\begin{aligned} [j_l, i_l] &= \min_i(\{[j,i] \,|\, [j,i] \in vertices(hull(\hat{edges}_l))\}), \\ [j_r, i_r] &= \max_i(\{[j,i] \,|\, [j,i] \in vertices(hull(\hat{edges}_r))\}) \end{aligned} \tag{B.14}$$

$$
\begin{aligned}
edges_l &= \{[j,i] \,|\, [j,i] \in \hat{edges}_l \wedge j \geq j_l\}, \\
edges_r &= \{[j,i] \,|\, [j,i] \in \hat{edges}_r \wedge j \geq j_r\}
\end{aligned} \tag{B.15}
$$

To refine the estimation of the radial boundaries, RANSAC [FB81] is applied to robustly regress a line of best fit, to determine the true radial boundaries of the plane. The RANSAC fit is performed for both the left $edges_l$ and right $edges_r$ set of edge points.

$$
\begin{aligned}
[M_l, residual_l] &= \text{RANSAC}_{\text{fit}}(edges_l) \\
[M_r, residual_r] &= \text{RANSAC}_{\text{fit}}(edges_r)
\end{aligned} \tag{B.16}
$$

Here, M are the model parameters and $residual$ is the fitting error. From the fitting, a linear representation of the radial boundaries is achieved as follows:

$$
\begin{aligned}
\hat{RB}_l &= \{M_l(min(edges_{l,i}), M_l(max(edges_{l,i})\} \\
\hat{RB}_r &= \{M_r(min(edges_{r,i}), M_r(max(edges_{r,i})\}
\end{aligned} \tag{B.17}
$$

Assuming that the centre of mass m splits the US plane symmetrically, the $\text{RANSAC}_{\text{fit}}$ producing the smallest residual error $[residual_l, residual_r]$ is chosen and re-projected across the vertical axis of symmetry defined by the line passing through m. This approach is based on the assumption that one side is likely to be worse effected by noise and discontinuities than the other. This is due to factors such as acoustic shadowing, which may result from orientation-related probe-tissue contact breaking. Using symmetry, the better-fit line is expected to be more reliable.

$$
RB_l = \begin{cases} \Phi - (\hat{RB}_r - \Phi), & \text{if } residual_r < residual_l \\ \hat{RB}_l, & \text{otherwise} \end{cases} \tag{B.18}
$$

$$RB_r = \begin{cases} \Phi + (\Phi - \hat{RB_l}), & \text{if } residual_r > residual_l \\ \hat{RB_r}, & \text{otherwise} \end{cases} \tag{B.19}$$

where, $\Phi = \{inner_{arc}, outer_{arc}\}$.

B.2.4 Annulus Sector Parameters Calculation

The origin of the annulus sector can be determined by finding the point of intersection of the radial boundaries. The end points of the radial boundaries are defined as $RB_{l,top}$, $RB_{l,bottom}$, $RB_{r,top}$, $RB_{r,bottom}$. The origin O is calculated by determining the coefficients a, b, c of the line equation, followed by the Cramer's rule:

$$\begin{aligned} a &= RB_{top}[1] - RB_{bottom}[1], \\ b &= RB_{bottom}[0] - RB_{top}[0], \\ c &= RB_{top}[0] \cdot RB_{bottom}[1] - RB_{bottom}[0] \cdot RB_{top}[1] \\ L &= [a, b, -c] \end{aligned} \tag{B.20}$$

$$\begin{aligned} D &= L_l[0] \cdot L_r[1] - L_l[1] \cdot L_r[0], \\ Dx &= L_l[2] \cdot L_r[1] - L_l[1] \cdot L_r[2], \\ Dy &= L_l[0] \cdot L_r[2] - L_l[2] \cdot L_r[0], \\ O &= \left[\frac{Dx}{D}, \frac{Dy}{D}\right] \end{aligned} \tag{B.21}$$

Once the origin O has been estimated, the angle θ between the intersecting lines can then be calculated as:

$$\theta = \arccos \frac{\overrightarrow{BA} \cdot \overrightarrow{BC}}{\|\overrightarrow{BA}\| \|\overrightarrow{BC}\|} \tag{B.22}$$

where, $A = RB_{l,bottom} - O$, $B = O$, $C = RB_{R,bottom} - O$.

From this, the inner and outer radius of the annulus sector can be estimated using the point of origin and the point of intersection of the inner and outer arcs along the x-coordinate of point m:

$$r_{inner} = \sqrt{||arc_{inner} - O||^2}$$
$$r_{outer} = \sqrt{||arc_{outer} - O||^2} \tag{B.23}$$

At this point, all relevant geometric information from the annulus sector has been processed. That is, the origin O, the angle θ, and the inner and outer radii $[r_{inner}, r_{outer}]$.

B.2.5 Scan line extraction and linearisation

Most US machines, particularly clinical ones, do not provide access to the original radio frequency (RF) data or plane information, often restricting these features to users. For example, RF data can be used to accurately process the image in such a way that the phase information is preserved. A primary motivation for the proposed method is the acquisition of scan lines, which offer the closest approximation to the original raw signals in the absence of RF data. The extraction of scan lines provides several advantages, such as enabling explicit processing along scan lines such as in [KWKN12b]. Another application of this would be accurate data augmentation, which is particularly useful for enhancing data diversity in neural network training. This section demonstrates how the method proposed above can be applied to linearise a convex US plane.

Using the estimated geometric parameters of the annulus sector, the US plane can be fanned over between the left and right radial boundaries to extract the start $[i_{inner}, j_{inner}]$ and end

$[i_{outer}, j_{outer}]$ of the scan line vectors as:

$$i_{inner} = r_{inner}\cos(\hat{\theta}) + O[1], j_{inner} = r_{inner}\sin(\hat{\theta}) + O[0],$$
$$i_{outer} = r_{outer}\cos(\hat{\theta}) + O[1], j_{outer} = r_{outer}\sin(\hat{\theta}) + O[0] \tag{B.24}$$

where, $\hat{\theta}$ is in the range $[90° - \theta/2, 90° + \theta/2]$ with *step* which can be defined by dividing θ by the length of the outer arc. Knowing θ and $[r_{inner}, r_{outer}]$, spline interpolation can be performed to extract subpixel representations of the scan line vectors. The number of samples for interpolation and the length of the scan lines (i.e. the linear image height), can be calculated by first determining the area of the annulus sector $Area = \theta \times (r^2_{outer} - r^2_{inner})/2$, then the $Height = r_{outer} - r_{inner}$ and $Width = Area/Height$. This produces a $Ratio = Height/Width$ so that the number of samples for interpolation can be determined using $step \times Ratio$. This ratio is key to maintaining the correct resolution and can also be used to control the computation cost (for example down-sampling). The lineralised image Ω is estimated as:

$$\Omega = [interpolate\left(I_{rgb}, \Gamma_{\hat{\theta}}\right), \forall\hat{\theta} \in \theta]$$
$$\Gamma_{\hat{\theta}} = [[i_{inner}, j_{inner}], [i_{outer}, j_{outer}]] \tag{B.25}$$
$$\tag{B.26}$$

where, $\Gamma_{\hat{\theta}}$ is the set of all beginning and end points for all scan lines. The pseudocode of the full linearisation process is shown in Algorithm 1.

Algorithm 1 Annulus Sector Linearisation

Input: Convex ultrasound image $I_{\mathbf{rgb}}$
Output: Linearized ultrasound image Ω
$I_g \leftarrow$ Convert I_{rgb} to grayscale
$I_b \leftarrow$ Binarize I_g using z-score for foreground detection
$I_c \leftarrow$ Extract largest connected component from I_b
$\hat{S} \leftarrow$ Compute negative cumulative sum of top half of I_c
$S \leftarrow$ Apply moving average on $\hat{S}$ to smooth
$m \leftarrow$ Calculate the average of the center of mass and the midpoint between minimas of S
$arc_{inner}, arc_{outer} \leftarrow$ Determine vertical intersections based on m in I_c
$edges_l, edges_r \leftarrow$ Determine radial boundary edges on I_c
$\hat{RB}_l, \hat{RB}_r \leftarrow$ Robust line fitting using RANSAC on $edges_l, edges_r$
$residual_l, residual_r \leftarrow$ Compute residuals for $\hat{RB}_l, \hat{RB}_r$
$RB_l, RB_r \leftarrow$ Apply symmetric correction based on geometric properties relative to $top, bottom$
$O \leftarrow$ Find intersection of RB_l and RB_r using Cramer's rule
$r_{inner}, r_{outer}, \theta \leftarrow$ Determine geometric parameters of annulus sector
$\Omega \leftarrow$ Perform spline interpolation using $r_{inner}, r_{outer}, \theta$

B.3 Experimental Setup and Results

B.3.1 Data

A controlled experiment, with minimal noise and corruption of the images, was carried out to validate the proposed approach on 30 images taken intraoperatively during brain surgery at Charing Cross Hospital, London, UK. The study had full local ethical approval from the HRA and Health and Care Research Wales (HCRW) authorities[1] A Canon i900 US machine (Canon Medical Systems, Japan) was used with an 8MHz i8MCX1 microconvex probe. The images are of size 960×1280. For the expansion of the evaluation to noisy public data, 50 selected images from POCUS [BWC+21b], DUKE [HLB+19], and CLUST [LBK+15] were processed and manually annotated. The images were annotated for all key points $[arc_{inner}, arc_{outer}, legs_{l,top}, legs_{l,bottom}, legs_{r,top}, legs_{r,bottom}, O]$ that were used as our ground truth.

[1]Study title - US-CNS: Multiparametric Advanced Ultrasound Imaging of the Central Nervous System Intraoperatively and Through Gaps in the Bone, IRAS project ID: 275556, Protocol number: 22CX7609, REC reference: 22/WA/0259, Sponsor: Research Governance and Integrity Team (RGIT).

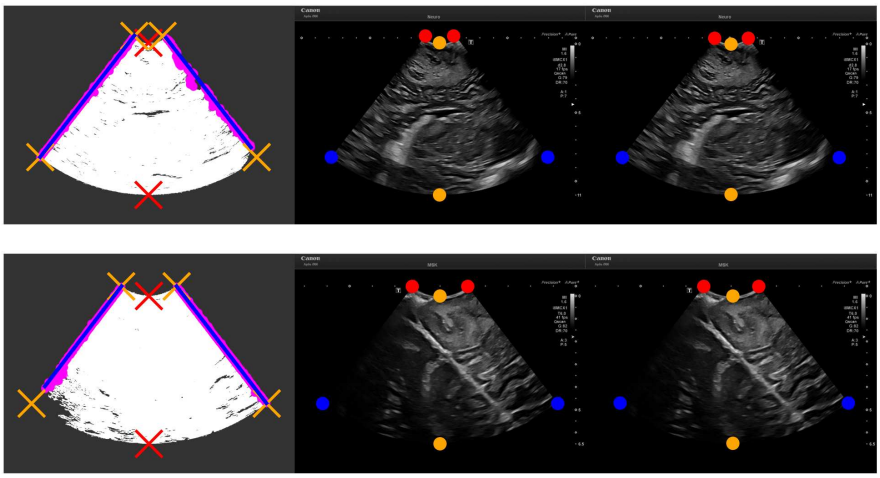


Figure B.5: Examples from the controlled experiment with the private data. The left images show the processing of the example images using the proposed method. The white plane is the output of the connected components process. The purple dots are the edge points of the mask, and the blue lines and orange crosses are the output of the RANSAC and the estimated radial boundaries. The middle and right images show the estimated and ground truth key points.

B.3.2 Key Point Detection Evaluation

B.3.2.1 Controlled Experiment

In this section, we evaluate the accuracy of our method in extracting the annulus sector information from private US images. For this purpose, the Mean Square Error (MSE) (in pixels) is calculated for each key point separately, as:

$$\text{MSE}^{kp} = \frac{1}{30}\sum_{f=1}^{30}\left((j_{gt}^{(f,kp)} - j_{pred}^{(f,kp)})^2 + (i_{gt}^{(f,kp)} - i_{pred}^{(f,kp)})^2\right) \tag{B.27}$$

where, kp is the key point index, f is the image index, $[gt, pred]$ denote the ground truth and predicted estimates, and i, j are the key point coordinates.

Across the 30 images, the following key point MSE were recorded: arc_{inner} = 3.05 (std 2.38), arc_{outer} = 3.24 (std 2.20), $legs_{l,top}$ = 4.41 (std 4.7), $legs_{l,bottom}$ = 4.50 (std 10.82), $legs_{r,top}$ = 5.90 (std 5.80), $legs_{r,bottom}$ = 6.51 (std 10.00), O = 3.72 (std 2.83) (all in pixels).

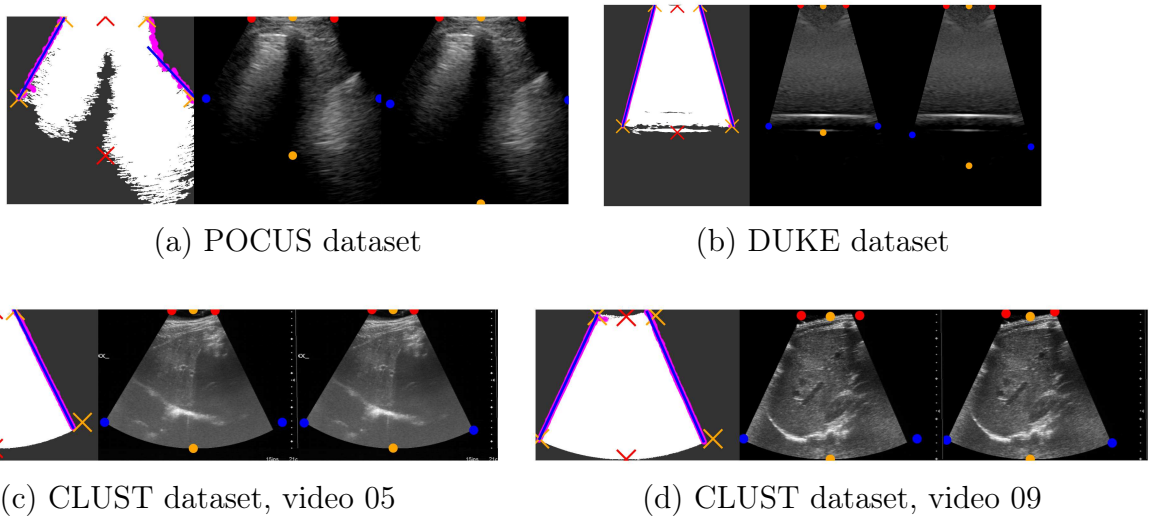
The error in the estimation of θ was also measured using the Mean Absolute Angular Difference

(MAAD) (in degrees) as:

$$\text{MAAD} = \frac{1}{30}\sum_{f=1}^{N}\left|\theta_{gt}^{f} - \theta_{pred}^{f}\right| \tag{B.28}$$

The recorded MAAD for our dataset was 0.30° (std 0.4°).

B.3.2.2 Public, Noisy Data



(a) POCUS dataset (b) DUKE dataset

(c) CLUST dataset, video 05 (d) CLUST dataset, video 09

Figure B.6: Example images from the public data experiment, one example from each public dataset used. The left image is the output of the method, the middle is the estimated key points and the right is the ground truth. All examples highlight the issue of cropping on the top of the US plane. The POCUS and DUKE examples show the issue of US plane corruption. Both CLUST examples suffer from slight rotation.

Public US images are rarely stored in a standardised format without corruption or alteration of the US plane. Often, images are cropped, rotated, warped, or noisy. In addition, they may be affected by other factors, such as capture from different devices and machines with varying settings. Examples of the data used in this evaluation are shown in Fig. B.6. This degree of irregularity in the data makes it challenging to utilise different data sources together in a collective way. The method proposed in this paper is an approach to alleviate this issue by being able to extract the underlying geometric properties of any US image. To evaluate the performance of our method on diverse data, images from multiple public datasets were annotated. Specifically, 18 images were taken from POCUS [BWC+21b], 17 from DUKE [HLB+19], and 15 from CLUST [LBK+15]. The selected images were chosen for their variability in the issues mentioned above.

Table B.1: MSE and MAAD results from the public data evaluation.

Dataset	**MSE** (O)		**MSE** (arc_{inner})		**MSE** (arc_{outer})		**MAAD** (°)	
	Mean	**Std**	**Mean**	**Std**	**Mean**	**Std**	**Mean**	**Std**
POCUS	23.73	43.38	4.17	7.19	23.59	43.04	1.98	2.47
DUKE	14.70	8.76	2.19	1.57	22.75	54.78	0.80	0.44
CLUST	20.85	4.93	8.48	2.15	9.06	2.05	2.97	0.024

There were two conditions used during data selection and annotation. Firstly, if the image was too challenging for human annotation, they would not be included in this test data. Secondly, the method is fundamentally designed for non-rotated or minimally rotated US-planes. Principally, this is because 2D US is not rotated in clinical practice. So, only non-rotated or slightly rotated images were chosen for annotation. This limitation mostly only concerns reconstructed 3D US - as 2D US (the primary concern of this paper and the more common US data) should not be stored rotated. In this case, slices are often poorly standardised and the US plane can vary significantly or even be absent due to cropping. The CLUST examples shown in Fig. B.6 are the cases where there is a slight rotation.

To handle images where the top of the US plane has been cropped - refer to all examples in Fig. B.6. Normally, the crop will result in the US plane starting from outside the image plane. So, if the top rows of I_c contain the binary mask (i.e. a vector of 1's at the top of the image), then the image is determined as not containing a concave profile. In this circumstance, rather than calculating the centre of mass, the middle point between the top left and right edges of the binary mask is used as the point m. The evaluation is carried out using the estimated and manually annotated geometric properties of the annulus sector: O, θ, arc_{inner}, and arc_{outer}. The average shape of the POCUS images was 571×571, DUKE 1345×960, and CLUST 510×686. The results are presented in Tab. B.1.

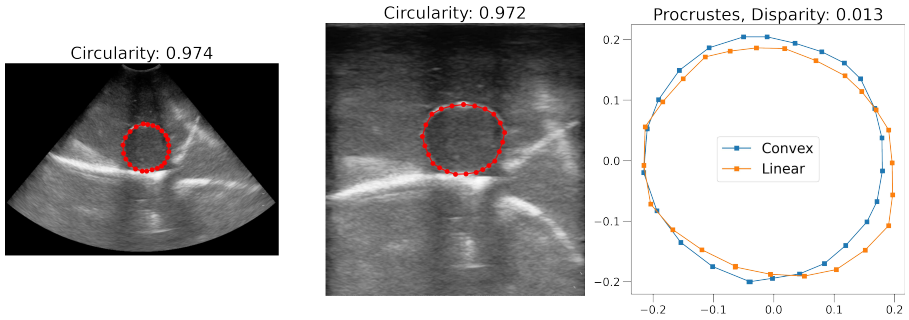


Figure B.7: Comparison of the circularity of the object in the original convex US plane compared to the linearised version. Quantitatively this comparison is measured using circularity and Proscrutes metrics.

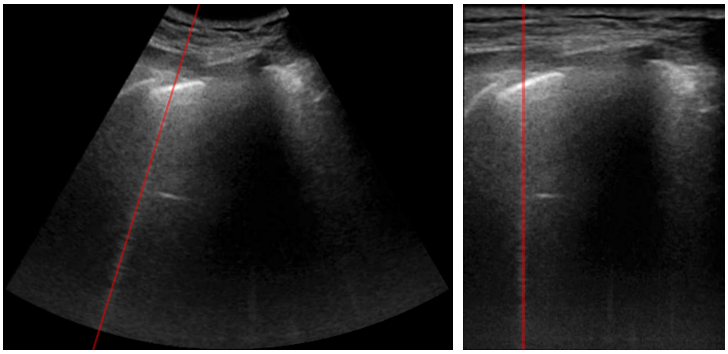


Figure B.8: Sample from POCUS dataset [BWC+21a] showing reverberation. The linearisation process correctly projects the reverberation as vertical.

B.3.3 Linearisation Study

B.3.3.1 Deformation

Experiments were conducted to assess the degree of deformation that occurs after a convex plane is linearised using our method. The purpose of this evaluation is to validate that the proposed method can correctly maintain the geometric features after augmentation. First, data was collected of a ping pong ball placed in water, using a Zonare Z.One Ultrasound System. The circularity of the ball was measured in both the original convex and extracted linear US images. For this purpose, in both images, the ping pong ball has been manually annotated and the circularity of both annotations is measured as:

$$\text{Circularity} = (4\pi\text{Area})/\text{Perimeter}^2 \tag{B.29}$$

As it is shown in Fig. B.7 (middle), the annotation on the extracted linear US image appears slightly stretched horizontally. However, this is somewhat expected as the linearisation will always deform the image to a certain degree. The quantitative results show that the circularity is almost equal in both images, approximately 0.97. Procrustes analysis [Gow75] is used to further compare the annotations by normalising them with respect to each other through translation and then isomorphic scaling and rotation. The result of this analysis is shown in Fig. B.7 (right). The recorded disparity, calculated as the sum of the squared pointwise differences is 0.013. Additionally, it can be observed that the contact-related acoustic shadow, the black columns on the left and right of the circle, is correctly linearised, with both columns projected as vertical lines.

Reverberations are common artefacts that appear as hyperechoic lines along scan lines, caused by repeated reflections between interfaces, including pleural line artefacts such as B-lines. To examine this, an image from the POCUS dataset [BWC+21a] is evaluated. The left image of Fig. B.8 shows a convex image with a red scan line which has been manually annotated by our clinical team as a reverberation. It can be assumed that the linearisation process should project reverberations as vertical columns. As can be observed, in the linearised image in Fig. B.8 (right), the reverberation is projected vertically, again along a scan line. This result shows that the linearisation process has preserved the true properties of the image, by maintaining the physics of the reverberation, without deformation.

B.3.3.2 Inversion

The aim of this section is to show that the linearisation process is not lossy and that the image augmentation process does not affect the underlying US information. The 30 intraoperative US images were linearised and then re-projected back to convex, using the extracted annulus sector geometry and $\hat{\theta}$. To measure the re-projection error, a direct MSE was calculated between the original and re-projected convex plane; across the 30 images the MSE was 0.0064 (std 0.0044) pixels.

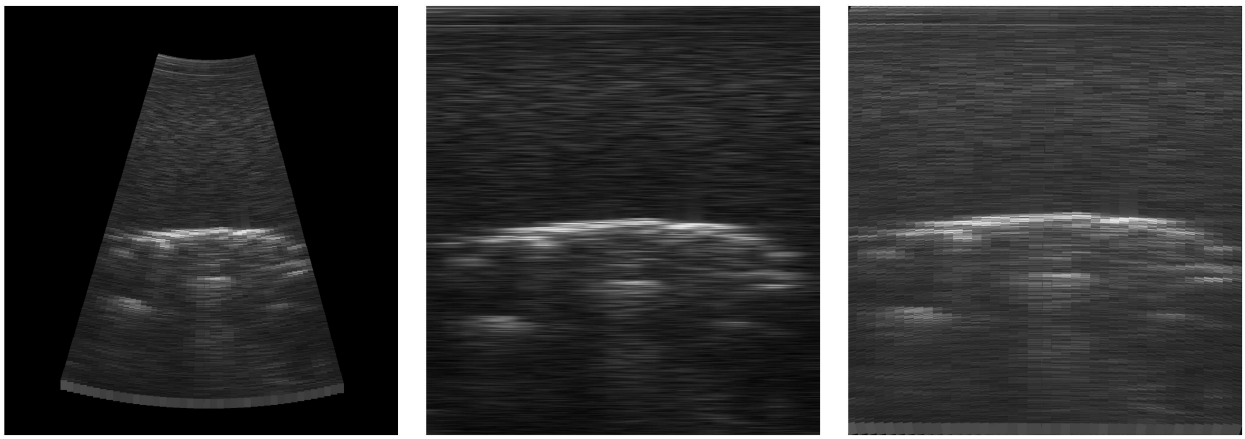


Figure B.9: Radio-frequency data validation. Left is the captured convex image. Middle is the ground truth linear image. Right is the linearised image estimated with our proposed method.

B.3.3.3 Radio-Frequency Data

The evaluation of the proposed linearisation method is extended with the use of RF data, which enables quantitative analysis of the linear output. The aim is to generate ground truth linear images using RF data, which can then be used to compare with the linearised images produced by the proposed method via structural similarity.

Access to raw RF data is not commonly available in clinical imaging systems, usually only in a limited number of research systems. B-mode US images provide a 2D visualisation of the RF data echo signals. A general representation of the detected echo $E^*(t)$ is:

$$E^*(t) = A(t)e^{j(2\pi ft+\varphi)} =$$
$$A(t)cos(2\pi ft + \varphi) + j_{rf}A(t)sin(2\pi ft + \varphi) = I + j_{rf}Q \tag{B.30}$$

where, $A(t)$ represents the signal envelope, f is the carrier frequency, φ is the signal phase, and I and Q represent the in-phase and quadrature sinusoids of the signal, respectively. For convex and phased array probes, each echo has a corresponding θ_{rf}, which relates to the orientation of the piezoelectric element. From the RF data, B-mode US images can be constructed by calculating the signal envelope, which then undergoes logarithmic compression and decimation.

A US slice was captured using a phased array US probe (P5-1S15-A6, Telemed, Lithuania), a digital acquisition board (DAQ) (ArtUS, Telemed, Lithuania) and a phantom made of Agarose

(2% v/v, Sigma-Aldrich) with a Polyvinil alcohol (PVA) circular inclusion (4% v/v, Sigma-Aldrich). Fig. B.9 shows the captured US data (left and middle) and the output of the proposed method (right).

Direct comparison of the intensities between the ground truth linear image from projected RF data and linearised images generated from convex B-mode is not feasible due to pixel discrepancy in the RF projection. Due to piezoelectric tilting θ_{rf}, corresponding convex and projected linear B-mode images exhibit variations in resolution. These variations lead to pixel-level discrepancies, which arise from differences in scan line spacing, overlap, and angular coverage. Moreover, since the data were acquired using a phased array, the convex image is generated through a projection process, in which the inner null region is simulated by prepending a zero vector. To mitigate the impact of these pixel-level differences, the evaluation of image linearisation is performed at the structural level using the multi-scale structural similarity metric (MS-SSIM) [WSB03].

$$\text{MS-SSIM} = [l_M] \prod_{g=1}^{M} [c_g][s_g] \tag{B.31}$$

where, l is the luminance, c is the contrast and s is the structural similarity between the two images, and g denotes the scale index during downscaling where $M = 5$ corresponds to the coarsest level. The result of this experiment was MS-SSIM equal to 0.69, which shows strong similarity between the ground truth linear and the proposed method's linear output. This discrepancy can be attributed towards the fundamental intensity differences on the pixel level between the convex and linear data. Through qualitative and quantitative analysis with ground truth, the results show that the proposed methodology is capable of accurately linearising convex US data.

B.4 Conclusion

This article proposes a new perspective and solution to alleviating the issue with online US data non-standardisation via harmonisation of data captured using different machines and data recording/saving methods. A method has been proposed to extract and represent arbitrary convex US planes, through estimation of its annulus sector parameters. This automatic estimation method is beneficial primarily because it enables standardisation of US datasets without needing additional manual annotation. Furthermore, through the estimation and processing of 2D convex planes using annulus sector properties, the scan lines can be extracted, and the US plane can be augmented, which could be used for neural network dataset supplementation. The proposed methodology is built on traditional image processing techniques and is designed to be robust to incomplete and corrupted presentations of US planes. Through evaluation of the accuracy of extracted key points, it can be concluded that our proposed method is capable of accurately finding the edges of the US plane and determining the parameters of the annulus sector. The evaluation is expanded to show that the annulus sector is capable of extracting the scan lines and projecting into a linear format, free of deformation.