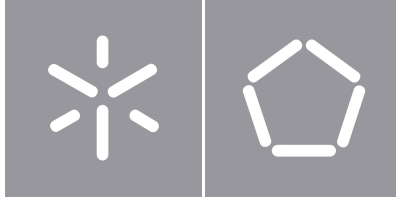


University of Minho
School of Engineering

Raquel Vitorino Lima

**Volume Reconstruction Algorithm for
Tumor Assessment from Freehand
Ultrasound Recordings in Mice**



University of Minho
School of Engineering

Raquel Vitorino Lima

**Volume Reconstruction Algorithm for
Tumor Assessment from Freehand
Ultrasound Recordings in Mice**

Master's Dissertation in Biomedical Engineering

Dissertation supervised by

Professor Doutor Jaime Francisco Cruz Fonseca
Doutor Sandro Filipe Monteiro Queirós

Copyright and Terms of Use for Third Party Work

This dissertation reports on academic work that can be used by third parties as long as the internationally accepted standards and good practices are respected concerning copyright and related rights.

This work can thereafter be used under the terms established in the license below.

Readers needing authorization conditions not provided for in the indicated licensing should contact the author through the RepositóriUM of the University of Minho.

License granted to users of this work:



CC BY-NC-ND

<https://creativecommons.org/licenses/by-nc-nd/4.0/>

“Believe you can and you’re halfway there.”

– Theodore Roosevelt

Acknowledgements

Ao meu orientador, Professor Doutor Jaime Fonseca, quero agradecer a disponibilidade e a orientação prestadas ao longo deste percurso. Ao meu coorientador, Doutor Sandro Queirós, agradeço pela dedicação e pelo acompanhamento constante, por me incentivar a sair da minha zona de conforto e por me desafiar de forma construtiva, contribuindo significativamente para o meu crescimento.

À Doutora Céline Gonçalves e à Doutora Alice Miranda, agradeço pela participação ativa nas aquisições realizadas para a criação do dataset desta tese e pela disponibilidade constante para me ajudar. Ao Doutor Fernando Veloso, agradeço todo o apoio no desenvolvimento do setup de validação utilizado neste trabalho.

Às tratadoras Susana e Cláudia, e aos investigadores Andreia Monteiro, Margarida Oliveira, Inês Castro e André Mira, agradeço a disponibilidade e todo o apoio prestado no âmbito do biotério.

Aos meus colegas e amigos de gabinete, Bruno, Marcos, João e Edgar, deixo um agradecimento especial por me terem acolhido desde o início, pela boa disposição diária e pelo apoio constante.

Aos meus amigos, não há palavras que expressem todo o amor, apoio e companheirismo que senti ao longo destes anos. À Mariana e à Nana, quero agradecer por serem as amigas desde o primeiro dia desta jornada, por estarem sempre presentes e por todos os momentos que tornaram este percurso mais leve e especial. Ao Jorge, por ser o amigo e colega de grupo em quem sempre pude confiar, por me desafiar e por me incentivar continuamente a fazer mais e melhor. À Inês, pela cumplicidade e pelo afeto, e por se ter tornado na flatmate que eu não sabia que precisava. Ao Hugo, pela escuta e amizade e à Bia, por me ensinar o valor da genuinidade e bondade. À Rita e à Andreia, pela alegria e otimismo constantes. À Cláudia, pelo apoio neste último ano e por me ajudar a ultrapassar tantas das minhas inseguranças. A todos vocês, e a muitos outros nomes que ficaram por mencionar, agradeço por me fazerem acreditar nas minhas capacidades quando eu própria duvidei. Foram cinco anos cheios de desafios e aprendizagens, que moldaram a pessoa que sou hoje e deixaram memórias que guardarei para sempre.

Aos meus pais e ao meu irmão, tudo o que sou a vocês o devo. Obrigada por serem um exemplo a seguir, por me apoiarem incondicionalmente e por me mostrarem que a resiliência e o esforço acabam sempre por ser recompensados. Esta conquista é tanto minha quanto vossa.

Statement of Integrity

I hereby declare having conducted this academic work with integrity.

I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration.

I further declare that I have fully acknowledged the Code of Ethical Conduct of the University of Minho.

University of Minho, Braga, january 2026

Raquel Vitorino Lima

Abstract

Accurate tumor volume quantification is crucial in mouse studies, yet existing three-dimensional (3D) imaging modalities such as micro-computed tomography and magnetic resonance imaging are slow, expensive, and require anesthesia. Ultrasound (US) provides a faster and more accessible alternative. However, in small-animal imaging, 3D US acquisition relies on mechanical scanning with a 2D probe, making it impractical for routine longitudinal studies and still requiring anesthesia. Freehand US could overcome these limitations, though its use is hindered by the challenge of estimating probe motion during acquisition.

In this context, this work explored two complementary approaches for reconstructing 3D volumes from freehand two-dimensional (2D) US in mice: a tracker-based method and a trackerless deep learning (DL) approach for pose estimation. In the first method, an Aurora motion tracking system (MTS) sensor was attached to the probe to record its pose over time, enabling volume reconstruction from the acquired sequence. This pipeline involved multiple stages, from sensor holder design and calibration to data acquisition, preprocessing, and volume reconstruction. For the trackerless approach, a DL model using EfficientNet-B1 architecture was adopted due to its efficiency and strong capacity to extract spatial features from US images. Multiple input/output network configurations and ensemble strategies were evaluated to improve pose prediction and assess the feasibility of this approach for freehand reconstruction.

This work demonstrated that tracker-based US reconstruction in mice is feasible, producing anatomically coherent volumes consistent with mechanical ground truth. Regarding the DL experiments, which were conducted on a publicly available dataset rather than on mice data that was still being acquired, the results highlighted how different model design choices affect prediction accuracy, with the ensemble strategies showing particularly promising improvements in the performance of the adopted network.

In summary, this thesis established a validated tracker-based framework for freehand US volume reconstruction in mice and explored DL as a potential trackerless alternative. The outcomes indicate that accurate volumetric imaging is achievable and that further optimization of the DL approach could allow fully trackerless reconstruction in future studies.

Keywords Deep learning, Freehand US reconstruction, Tracker-based and trackerless pose estimation.

Resumo

A quantificação precisa do volume tumoral é essencial em estudos com ratinhos, mas as modalidades de imagiologia tridimensional (3D), como a micro-tomografia computadorizada e a ressonância magnética, são demoradas, dispendiosas e requerem anestesia. O ultrassom (US) constitui uma alternativa mais rápida e acessível. Contudo, a aquisição 3D em ratinhos depende de sistemas de varrimento mecânico, os quais são pouco práticos para estudos longitudinais e exigem anestesia. A aquisição livre de US pode ultrapassar estas limitações, embora seja difícil estimar o movimento da sonda durante a aquisição.

Neste contexto, esta dissertação explorou duas abordagens para a reconstrução a partir de aquisições livres de US: um método baseado em sistemas de *tracking* e uma abordagem *trackerless* recorrendo a *DL* para a estimativa da pose. No primeiro método, um sensor de *tracking* foi acoplado à sonda para registar a sua pose ao longo da aquisição, permitindo a reconstrução a partir da sequência adquirida. Esta metodologia envolveu várias etapas, desde a calibração e aquisição até à reconstrução volumétrica. Na abordagem *trackerless*, foi utilizado um modelo de aprendizagem baseado na arquitetura *EfficientNet-B1*, devido à sua eficiência e capacidade de extrair características espaciais das imagens de US. Foram testadas diferentes configurações da rede e estratégias de *ensemble* para melhorar a estimativa da pose e avaliar a viabilidade desta abordagem para a reconstrução de US.

Os resultados demonstraram que a reconstrução baseada em *tracking* é viável, produzindo volumes consistentes com o volume de referência por varrimento mecânico. A pipeline completa de aquisição e calibração foi implementada e validada com sucesso, resultando em reconstruções coerentes. Relativamente à abordagem *trackerless*, os resultados evidenciaram como as diferentes escolhas de configurações do modelo influenciam a precisão, sendo que as estratégias de *ensemble* revelaram melhorias promissoras.

Em suma, esta dissertação estabeleceu uma *framework* válida para a reconstrução volumétrica de aquisições livres de US em ratinhos e explorou técnicas de *DL* como alternativa *trackerless*. Os resultados indicam que a imagiologia volumétrica precisa é alcançável, e que otimizações futuras da metodologia de *DL* poderão permitir reconstrução totalmente *trackerless*.

Palavras-chave *Deep learning*, Reconstrução volumétrica, Métodos com e sem *tracking*.

Contents

1	Introduction	1
1.1	Context	1
1.2	Imaging modalities in mouse models	2
1.2.1	Micro-computed tomography	2
1.2.2	Magnetic resonance imaging	3
1.2.3	Ultrasound imaging	4
1.3	Freehand volume reconstruction	5
1.4	Motivation	7
1.5	Aims and contributions	8
1.6	Outline	9
2	State-of-the-art	10
2.1	Tracker-based vs trackerless estimation	11
2.2	Volume reconstruction	12
2.2.1	Pixel-based methods	12
2.2.2	Voxel-based methods	13
2.2.3	Function-based methods	13
2.3	Trackerless volume reconstruction	14
2.3.1	Classical approaches	14
2.3.2	Deep learning approaches	18
2.4	Evaluation metrics	29
2.4.1	Frame-to-frame error	29
2.4.2	Volume error	31
3	Tracker-based pose estimation	36
3.1	Methods	36

3.1.1	System and hardware	37
3.1.2	Calibration procedures	39
3.1.3	Data acquisition and preprocessing	44
3.1.4	Volume reconstruction	46
3.2	Experiments	46
3.2.1	Tracker-ultrasound validation	47
3.2.2	Volume reconstruction validation	51
3.3	Results and discussion	53
3.3.1	Tracker-ultrasound validation	53
3.3.2	Volume reconstruction validation	58
3.4	Conclusion	60
4	Trackerless pose estimation	61
4.1	Methods	61
4.1.1	Overview	62
4.1.2	Model architecture	62
4.1.3	Input variants	63
4.1.4	Output variants	64
4.1.5	Ensembles	68
4.1.6	Loss functions	69
4.1.7	Implementations details	70
4.2	Results and discussion	70
4.2.1	Dataset	70
4.2.2	Evaluation metrics	71
4.2.3	Network performance	73
4.2.4	Conclusion	79
5	Conclusion and future developments	80
	Appendices	91
A	Design of the sensor holder	91
B	CAD schematics for tracker–US validation	93

List of Figures

1	Schematic representation of a micro-CT system	3
2	MRI system for preclinical studies in small animals	4
3	Illustration of ultrasound (US) imaging acquisition in mice	5
4	Schematic representation of the setup for freehand two-dimensional (2D) US imaging . .	6
5	Schematic overview of three-dimensional (3D) volume reconstruction from 2D freehand US acquisitions	10
6	Available motion tracking systems (MTSs)	11
7	Overview of the speckle decorrelation framework	15
8	Fundamental concept of elevational speckle decorrelation	15
9	Illustration of the rigid motion estimation process using local 3D translation measurements	16
10	Illustration of the approach by Chang <i>et al.</i>	16
11	Example movement types encountered during a freehand scan	17
12	Freehand volume reconstruction framework proposed by Tetrel <i>et al.</i>	18
13	General workflow of convolutional neural network (CNN)-based methods for pose estimation	19
14	CNN architecture proposed by Miura <i>et al.</i>	20
15	Workflow of the approach proposed by Wein <i>et al.</i>	21
16	deep contextual learning network (DCL-Net) architecture proposed by Guo <i>et al.</i>	22
17	Modified contrastive-based DCL-Net architecture proposed by Guo <i>et al.</i>	22
18	Multi-task learning approach proposed by Li <i>et al.</i>	23
19	3D partial convolution network proposed by Dong <i>et al.</i> for freehand volume reconstruction	24
20	Online learning framework proposed by Luo <i>et al.</i>	25
21	Improved online learning framework proposed by Luo <i>et al.</i>	26
22	Workflow proposed by Leblanc <i>et al.</i>	27
23	Overview of the method proposed by Dou <i>et al.</i>	28

24	Workflow proposed by Li <i>et al.</i>	28
25	Overview of the proposed MTS-based framework	37
26	Data flow between the core components of the Aurora motion tracking system	38
27	Components used during the calibration procedures	41
28	Lateral views of the phantom illustrating the printed reference points employed in the registration step	42
29	Coordinate systems of the components involved in the calibration process	43
30	Data acquisition and preprocessing workflow	44
31	Validation setup showing the validation support and the acrylic assembly	47
32	Illustration of the experimental setups used in the validation	48
33	Spatial visualization of positional and orientation precision for the experimental setup	55
34	Spatial visualization of positional and orientation accuracy for the experimental setup	56
35	Example of the 3D distribution of the segmented wire points and corresponding best fit lines in the parallel setup	58
36	Qualitative comparison of kidney segmentations from MTS-based reconstruction and mechanical scanning	60
37	Example of tumor acquisition and reconstruction.	60
38	Overview of the EfficientNet-B1 architecture	63
39	Illustration of the paired-frame and window-based input strategies for pose estimation	64
40	Best inference cases for patient 0 using the 10-frame reference model	77
41	Worst inference cases for patient 0 using the 10-frame reference model	77
42	Technical drawing of the MX250 probe	91
43	Technical drawing of the holder designed to secure the sensor to the probe	92
44	Technical drawing of the MX250 probe showing the sensor attached to the holder	92
45	Sensor holder design in SolidWorks and its real 3D-printed implementation attached to the MX250 probe	92
46	Technical drawings of the validation support structure	93
47	Technical drawings of the custom validation setup developed for this study	94
48	CAD views of the validation model showing the front, isometric, and side perspectives.	94
49	Different views of the validation model inside the water container	94

50	Internal wire configuration of the validation model	94
----	---	----

List of Tables

1	Precision and accuracy metrics for the system under the parallel and perpendicular setups.	54
2	mean absolute distance (MAD) across calibration sets – combined parallel and perpendicular setups	57
3	Angular deviation between adjacent positions	58
4	Comparison of motion tracking system (MTS)-based and mechanical-scan volumes . . .	59
5	Performance for different input/output configurations	74
6	Performance of all ensemble configurations obtained for the best-performing model. . .	78
7	Performance comparison for different Pearson correlation weights combined with MSE loss	79

Acronyms

1D one-dimensional.

2D two-dimensional.

3D three-dimensional.

4D four-dimensional.

ADL adversarial learning.

ASSD average symmetric surface distance.

CAD computer aided-design.

CNN convolutional neural network.

ConvLSTM convolutional long short-term memory.

CT computed tomography.

DCL-Net deep contextual learning network.

DL deep learning.

DoF degree of freedom.

FBM function-based method.

FC fully connected.

FCC frame-level contextual consistency.

FD final drift.

FDR final drift rate.

FDS fully developed speckle.

FG field generator.

FN false negatives.

FP false positives.

fps frames per second.

GAN generative adversarial network.

GAP global average pooling.

GAS global adversarial shape prior.

GCE global corner reconstruction error.

GMS gradient magnitude similarity.

GMSD gradient magnitude similarity deviation.

GPE global pixel reconstruction error.

GT ground truth.

HD Hausdorff distance.

HD95 95th percentile Hausdorff distance.

IMU inertial measurement unit.

LCE local corner reconstruction error.

LPE local pixel reconstruction error.

LSTM long short-term memory.

MAD mean absolute distance.

MAE mean absolute error.

MBConv mobile inverted bottleneck convolution.

micro-CT micro-computed tomography.

MRI magnetic resonance imaging.

MSE mean squared error.

MTS motion tracking system.

PBM pixel-based method.

PIQE perception-based image quality evaluator.

PLUS Public Software Library for Ultrasound.

PNN pixel nearest neighbor.

PSC path-level similarity constraint.

RBF radial basis function.

ResNet residual network.

RF radio frequency.

RiIG Rician Inverse Gaussian.

RMSE root mean squared error.

RNN recurrent neural network.

ROI region of interest.

RVD relative volume difference.

SCU system control unit.

SIU sensor interface unit.

SND smooth neighborhood descriptor.

SSIM structural similarity index metric.

SSL self-supervised learning.

TP true positives.

TUS-REC Trackerless 3D Freehand Ultrasound Reconstruction.

US ultrasound.

VBM voxel-based method.

VNN voxel nearest neighbor.

XML extensible markup language.

Chapter 1

Introduction

1.1 Context

When it comes to advancing science and understanding the evolution of human diseases, mouse models are the most widely used experimental models in biomedical research. Humans and mice are genetically and physiologically similar, with around 90% of their genomes organized in the same order, known as conserved syntenic regions [1]. This means that the order and orientation of genes on chromosomes are largely the same in the two species, allowing direct comparisons of genetic function and regulation. As a result, it is possible to study and investigate diseases, develop medicines, and evaluate therapies, thus contributing to the understanding of biological and pathological mechanisms [2, 3].

According to the World Health Organization, cancer remains one of the leading causes of mortality worldwide, with an estimated 20 million new diagnoses and 9.7 million deaths in 2022 [4]. Approximately 53.5 million individuals were alive within five years of receiving a cancer diagnosis. Around 20% of people will face a cancer diagnosis during their lifetime, and about 1 in 9 men and 1 in 12 women succumb to the disease. By 2050, the number of new cancer cases is anticipated to exceed 35 million, marking a 77% increase from the projected 20 million cases in 2022. When looking at the overall impact, countries with High Human Development Index are expected to experience the largest absolute rise in incidence, with an estimated 4.8 million additional cases in 2050 compared to 2022 [4].

To improve knowledge and develop more effective cancer treatments, extensive research has been carried out in this field using animal models. Approximately 95% of these studies involve subcutaneous injection of cancer cells into mice, enabling the study and monitoring of tumor growth and its response to administered drugs [3]. Analyzing the evolution of tumor size, texture, and boundaries is thus a critical step in evaluating the success of such research, as it offers valuable insights into how the model responds to the treatments given.

1.2 Imaging modalities in mouse models

In recent years, advanced imaging techniques such as micro-computed tomography (micro-CT) [5, 6], magnetic resonance imaging (MRI) [7, 8] and ultrasound (US) [9, 10] have facilitated non-invasive monitoring of tumor progression, offering precise measurements of tumor volume, vascularization, and tissue heterogeneity. These advancements have significantly impacted cancer research, enabling the discovery of innovative treatments and prevention strategies, improving the understanding of cancer biology, and supporting the identification of biomarkers and molecular subtypes [11].

1.2.1 Micro-computed tomography

Micro-CT, also known as micro X-ray computed tomography, is a high-resolution imaging method that has shown significant efficacy in research involving small animals. This imaging technique captures a large number of two-dimensional (2D) X-ray images from various perspectives and employs computerized processes to reconstruct them into a multidimensional model [5, 7, 12–14].

Similar to clinical computed tomography (CT), micro-CT scanners operate on the same fundamental physical principles but are specifically designed for imaging at a higher resolution. Initially, an X-ray source generates a cone-shaped beam directed at the target tissue. On the opposite side, an X-ray detector with a scintillation layer captures the energy absorbed by the X-rays passing through the sample. Due to the distinct attenuation coefficients of various tissues, the system is able to differentiate between structures within the tissue, enabling detailed visualization of anatomical features [5, 12]. However, this method involves radiation exposure and requires the subject to be anesthetized, as the resolution directly depends on how still the subject remains [12].

Following the acquisition of the 2D projections, they are reconstructed into three-dimensional (3D) images through mathematical algorithms. The most commonly employed approach is filtered back projection, which uses a convolution kernel to minimize blurring before back-projecting the data into the object space [5, 12]. While filtered back projection remains the standard method, iterative techniques such as the algebraic reconstruction technique provide superior results at the cost of efficiency. This approach improves the reconstructed image by iteratively analyzing simulated projections against the collected data, making them particularly effective when dealing with incomplete data. Figure 1 presents a schematic representation of a micro-CT system.

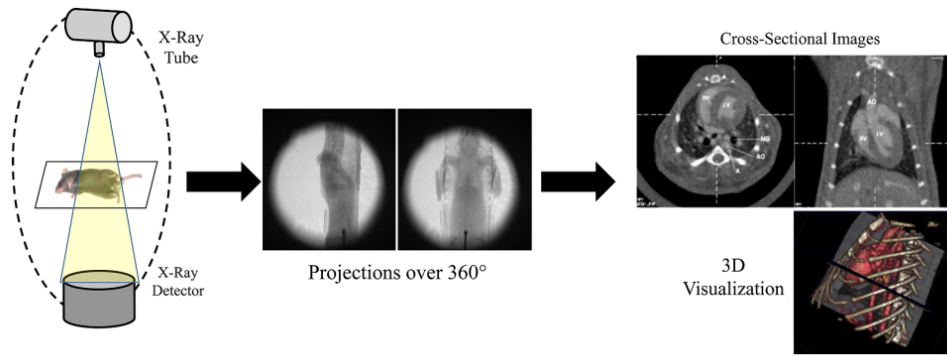


Figure 1: Schematic representation of a micro-CT system. From [5].

1.2.2 Magnetic resonance imaging

MRI is a non-invasive, radiation-free technique with a high signal-to-noise ratio and excellent spatial and temporal resolution [7, 15]. This method exploits the magnetic properties of hydrogen nuclei dispersed within the tissue being analyzed. Under normal conditions, the hydrogen atoms in tissue are randomly oriented. However, when exposed to an external magnetic field, these atoms align with the field's direction, moving away from their equilibrium state. Once the radio frequency (RF) excitation pulse is turned off, the nuclei relax back toward thermal equilibrium within the static magnetic field, releasing energy in the form of RF signals that are detected by receiver coils and converted into digital signals [16, 17].

MRI supports various types of sequences and imaging techniques, allowing the generation of images with diverse characteristics to evaluate functional parameters such as perfusion, water diffusion, and cellular metabolism, among others [15]. One of the most widely used methods in studying tumor growth in mice is T2-weighted imaging, which is sensitive to the specific properties of water diffusion. This technique facilitates the construction of cross-sectional views of tumor growth and assessment of therapeutic response [7].

As in micro-CT, live animal studies require mice to be anesthetized to ensure immobilization during image acquisition. Taking these factors into account, a MRI setup (Figure 2) usually includes an anesthesia system, a motion-restriction device, and a platform resembling a bed to position the subject [18]. To generate the magnetic field, RF coils are incorporated into the system.

Although MRI provides high-resolution imaging, it has limitations. The integration of RF coils into the system is technically complex and can compromise sensitivity. Additionally, the use of anesthesia subjects mice to increased stress, which can influence results and complicate post-procedure monitoring [18, 19].



Figure 2: MRI system for preclinical studies in small animals. From [18].

1.2.3 Ultrasound imaging

US is a non-invasive, quick, and relatively low-cost imaging modality compared to other techniques, enabling real-time visualization of tumors without the use of ionizing radiation [10, 20, 21].

US imaging is based on the interaction between the waves produced by a piezoelectric system and the target tissues. When a US wave penetrates the body, several interactions can occur, including reflection, refraction, dispersion, and attenuation. Reflection happens when US waves encounter tissues with different acoustic impedances, causing some waves to reflect back toward the probe. These reflected waves are then converted by the US system into electrical signals to create an image. Wave dispersion or scattering occurs when waves encounter structures, known as diffuse reflectors, that have a different acoustic impedance than the surrounding tissue and a shorter wavelength than the incident wave. This phenomenon causes multiple echoes propagating from many tiny structures to meet and interact with each other, leading to speckling in the images. Refraction occurs when US waves pass between tissues and are transmitted at an angle different from their original propagation path [22]. Figure 3 illustrates the imaging acquisition process in mouse models.

In US imaging, transducers can be classified according to their dimensionality, ranging from one-dimensional (1D) to four-dimensional (4D). For the analysis of tumors and pathologies, 3D probes would be the most suitable, as they can capture complete volumes with high spatial resolution [10]. However, 3D and 4D transducers are not available for imaging in mice, which presents a challenge in monitoring tumor growth.

Conversely, conventional 2D US scans are widely used in research due to their accessibility. However, they do not provide a complete view of anatomical structures, which significantly limits volume measurement and tissue identification [24], resulting in much lower accuracy in tumor assessment.

Several methods have been proposed to accurately create 3D US volume data, which can be classified into three different categories: dedicated 3D probes, mechanical scanning, and freehand scanning [24,

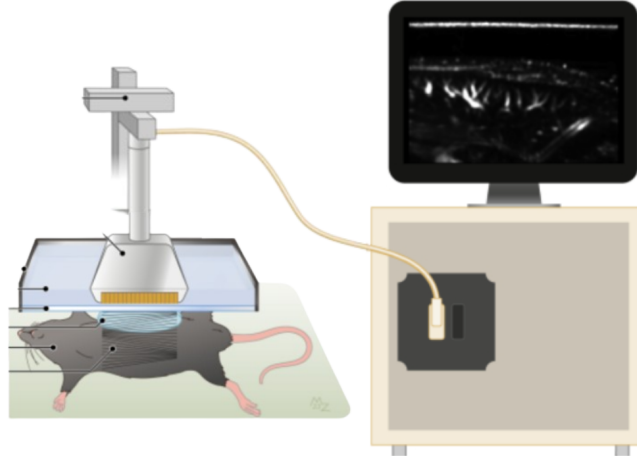


Figure 3: Illustration of US imaging acquisition in mice. From [23].

[25]. Dedicated 3D probes can offer real-time 3D volume data by using an advanced mechanism to scan a predetermined region of interest (ROI). Nevertheless, as mentioned above, this type of transducer is not available for imaging in mice. Mechanical scanning typically involves a stepping motor and a scanning head, which move and rotate the 2D transducer while simultaneously recording position and orientation information. However, this method has a limited scanning range [25] and requires the mice to be under the effect of anesthesia. Freehand scanning is a relatively low-cost and practical method for longitudinal studies, as it involves capturing images with the 2D probe being moved freely by the user, without the need for mechanical support. In this way, it allows the operator to acquire data from different angles and directions, depending on the desired visualization of the structure. The acquired data is then processed and transformed using specialized software that estimates the unknown spatial positions of each frame, enabling the reconstruction of the 3D volume. Given its advantages and adequateness for mice studies, this dissertation will focus on the latter. The following subsection details the volume reconstruction process using this technique.

1.3 Freehand volume reconstruction

Freehand volume reconstruction enables the transformation of a set of frame recordings captured by a 2D transducer into comprehensive 3D volume data.

The conventional approach to achieve this transformation typically requires tracking the 2D probe to obtain its position and orientation during image acquisition. Numerous methods have been proposed to estimate these parameters during freehand US imaging, including techniques involving position-sensing systems such as optical, acoustic, or electromagnetic tracking devices [26, 27]. This system fundamentally

consists of three components: a conventional 2D US scanner, a MTS device, and a workstation with specialized software for data collection, synchronization of signals from the US and the tracker's coordinate system, and 3D volume reconstruction [21], as shown in Figure 4. While these systems enable highly accurate reconstruction, MTSs are often expensive and may not be convenient for the user to handle the probe effectively during image acquisition.

Over the years, research has focused on developing more efficient methods for volumetric reconstruction that do not require a position sensor. To achieve this, it is assumed that the transformation occurring between successive frames is a rigid transformation, meaning that there is no change in the size or shape of the tissue being analyzed, and only translation and rotation occurred between the scans. Therefore, to estimate the transformation between each frame pair, it is necessary to calculate the six degrees of freedom (DoFs): three for translation along the X, Y, and Z axes, and three for rotation around these axes.

A reconstruction approach based on the analysis of the images obtained was initially proposed, aiming to recognize patterns such as speckle patterns. When a US wave interacts with the target tissue, scattering can occur, as explained above, causing a random distribution of intensities and resulting in a granular pattern known as speckle. By analyzing two consecutive frames, it is possible to calculate how the two images are related using speckle decorrelation algorithms. This process is repeated for the entire set of images, ultimately yielding the estimated transformation parameters for every frame. However, this approach presents challenges as the speckle pattern is rarely fully developed, and the actual content of the two images can differ. This discrepancy complicates the estimation of elevational displacement, limiting the accuracy of position and orientation measurements [26, 29].

In recent years, with advances in technology, approaches based on deep learning (DL) have been increasingly used to estimate how the obtained frames can be combined, employing convolutional neural

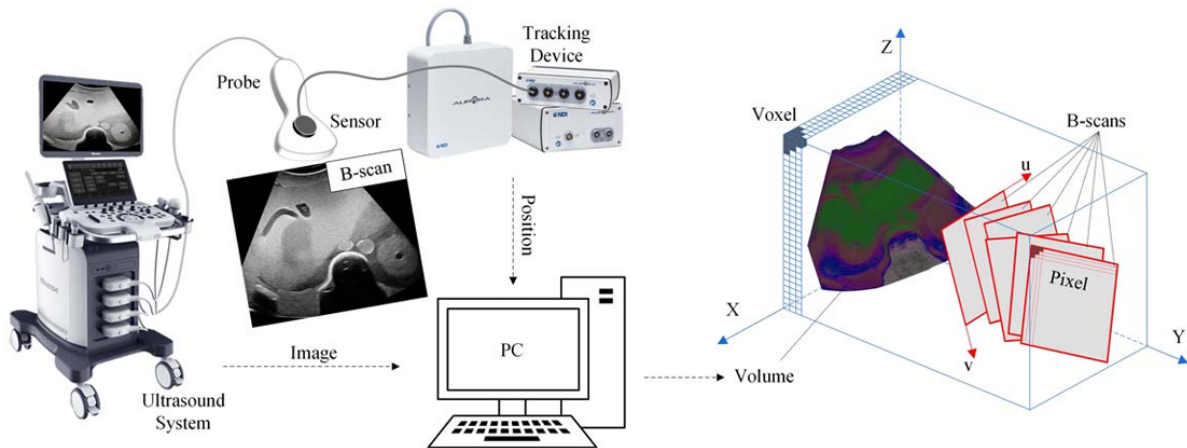


Figure 4: Schematic representation of the setup for freehand 2D US imaging. From [28].

networks (CNNs), long short-term memory (LSTM) networks, and residual networks (ResNets), among others [30, 31]. These approaches demonstrate enhanced efficiency and improved results compared to conventional methods, as they can learn meaningful image representations with expanded feature learning capabilities in an end-to-end fashion [31, 32].

After estimating the transformations required to align each frame, the 3D reconstruction can be performed using various interpolation techniques, including voxel-based methods (VBMs), pixel-based methods (PBMs) or function-based methods (FBMs).

1.4 Motivation

Cancer is one of today’s major humanitarian health crises, affecting millions of people around the world [4]. The high mortality rate and the increasing number of cases have led to intensive research efforts to understand and treat this pathology more effectively.

In the field of oncology research, particularly in studies involving mice, precise volumetric analysis is a critical step in the evaluation of tumors and, consequently, their response to treatment. With the advance of technology, various imaging modalities, such as MRI, micro-CT, and US, can be used to study tumors and their progression.

Currently, the methods used to analyze and monitor tumor size in mice include: (1) using calipers to estimate tumor volume based on two orthogonal measurements, but applicable to subcutaneous tumors only; (2) employing 2D US imaging to approximate tumor volumes from measurements of tumor length, width, and depth obtained from two orthogonal imaging planes; and (3) utilizing 3D US imaging combined with segmentation techniques to calculate accurate tumor volumes. It is important to note that the first two methods provide only approximate tumor volume estimates, which may lead researchers to incorrect conclusions, whereas the latter enables precise reconstruction, but requires the mouse to be anesthetized and involves the use of a motorized probe.

Over the years, US imaging has become increasingly popular due to its cost-effectiveness, accessibility, and radiation-free nature. Given these advantages, as well as the limitations of 3D probes and mechanical US systems, various techniques have been proposed to reconstruct 3D volumes from a sequence of 2D US images, relying on the tracking or estimation of the probe’s pose for accurate reconstruction.

MTSs have been considered a promising solution, as they provide reliable and accurate pose information for freehand reconstruction. However, they also introduce practical constraints. Optical systems require maintaining the camera’s field of view, while electromagnetic systems are sensitive to magnetic

disturbances [29]. In addition, they rely on specialized hardware, which increases both system cost and complexity.

Recently, rapid advances in DL methods have enabled the development of approaches that not only align the sequence of frames with high precision but also eliminate the need for MTSs. Despite these advances, none of these methods has yet been applied to freehand US reconstruction in animal imaging.

This thesis addresses these gaps by developing and validating a MTS-based pose estimation framework for freehand US reconstruction in small animal imaging, while also exploring DL-based pose estimation on a publicly available human forearm US dataset as a potential trackerless alternative.

1.5 Aims and contributions

The primary aim of this thesis is to investigate and evaluate two methodologies for freehand US reconstruction imaging: (1) reconstruction based on an external MTS; and (2) reconstruction based on DL-based pose estimation as a potential trackerless alternative. The performance of each approach is analyzed independently to assess its feasibility, accuracy, and potential suitability for preclinical tumor monitoring. The contributions of the present dissertation are as follows:

1. Reviewed the state-of-the-art in freehand US volume reconstruction, including both MTS-based and DL-based pose estimation methods;
2. Implemented an experimental setup for acquiring freehand US recordings in mice with associated spatial information using an electromagnetic MTS;
3. Assisted in the acquisition and curation of a dataset of freehand US recordings in mice;
4. Developed and validated a complete pipeline for MTS-based freehand US reconstruction, including tracking calibration, synchronization, and volume reconstruction;
5. Explored and implemented a DL-based approach to estimate probe motion directly from the analysis of US pixel data;
6. Evaluated the performance of multiple DL network configurations, including different input strategies, output representations, window sizes, and ensemble methods.

1.6 Outline

In the current chapter, an overview of the challenge addressed in this thesis was given, specifically the development of a method to monitor and measure tumors in research settings. Initially, global cancer statistics were introduced, emphasizing the importance of accurate evaluation methods. Various imaging modalities were then explored, highlighting US as the most accessible, cost-effective, real-time, and radiation-free option. Thereafter, a 2D US scanning technique, known as freehand US, was presented as a promising approach for this research. Finally, the thesis' motivation and objectives were outlined.

The second chapter reviews the state-of-the-art, detailing the main stages of volume reconstruction. It first examines the different methods for acquiring/estimating the probe's rotational and translational parameters across the frame sequence. Next, it presents the primary interpolation techniques used for volume reconstruction once the pose parameters have been determined. Then, trackerless parameter estimation methods, including classical and DL approaches, are analyzed in detail. The chapter concludes by summarizing the evaluation metrics used to assess model performance.

In the third chapter, the MTS-based pose estimation pipeline is presented in detail. The chapter begins by describing the experimental setup, including the US system, the electromagnetic tracking hardware, and the custom probe sensor holder designed to attach the sensor to the probe during data acquisition. This is followed by the calibration procedures required to align the coordinate systems of the tracker, probe, and imaging systems. The subsequent section details the acquisition protocol, data preprocessing steps, and the methods used for freehand volume reconstruction. The chapter then describes the experiments performed to validate both the tracking accuracy and the reconstruction quality. Finally, the results for each validation stage are presented and discussed.

In the fourth chapter, trackerless pose estimation based on DL is explored. Since the acquisition of the mice dataset was still ongoing and insufficient for training DL models, the proposed methods were developed and evaluated using a publicly available US dataset. The chapter begins with an overview of the proposed approach, followed by a description of the model architecture used in this work. Different input configurations and output representations are then presented, along with ensemble strategies evaluated to improve pose prediction. Finally, the chapter describes the experiments conducted to assess the performance of the different approaches, and presents the corresponding results and discussion.

The final chapter presents the conclusions of this work and discusses possible future developments.

Chapter 2

State-of-the-art

In the present day, US has gained increasing popularity due its ability to provide safe, cost-effective, and real-time imaging. In biomedical research using mouse models, 2D US is most commonly used, as it allows researchers to obtain detailed anatomical images. However, it is limited to the visualization of 2D slices of inherently 3D structures, making it difficult to perform accurate analyses and obtain a complete view of the ROI.

Although never applied before in the context of animal models, techniques such as freehand volume reconstruction are increasingly being used to generate volumes from non-parallel 2D image captures, effectively addressing the aforementioned challenge. As shown in Figure 5, the process of reconstructing a 3D volume from an acquired sequence of 2D US images can be subdivided into two main stages: (1) the estimation of how the images align with each other; and (2) the reconstruction of the 3D volume based on available images and respective alignment parameters obtained in the previous stage.

For frame alignment, two approaches can be utilized: a tracker-based method, where a MTS is directly attached to the probe, or a trackerless method, where the displacement between frames is estimated by analyzing the image content, using classical algorithms or DL techniques. Once the pose parameters are

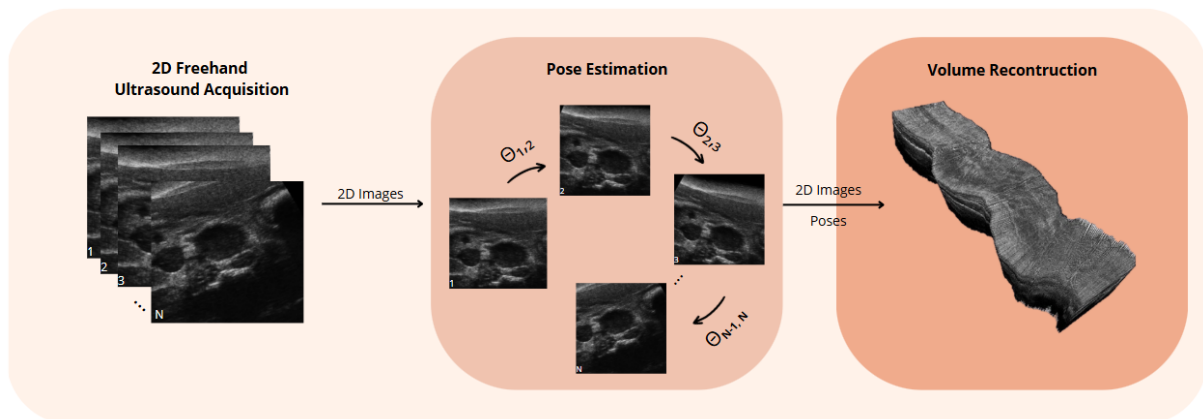


Figure 5: Schematic overview of 3D volume reconstruction from 2D freehand US acquisitions.

estimated, interpolation techniques are applied to create the reconstructed volume by representing each 2D frame within its corresponding 3D location.

2.1 Tracker-based vs trackerless estimation

To reconstruct volumes from a sequence of frames, it is first necessary to determine how the frames should be aligned, i.e. the frames' relative poses. There are two primary methods for estimating these transformation parameters: tracker-based systems and trackerless approaches.

In tracker-based approaches, the position and orientation of the probe are provided directly by MTS, which offers six DoFs. Various motion tracking technologies can be employed, with the most common being electromagnetic, optical, acoustic sensors, and inertial measurement units (IMUs) [27]. This methodology requires precise temporal and spatial calibration to synchronize the probe's image data with the motion tracking positioning data, ensuring accurate volume reconstruction. Figure 6 illustrates various MTSs available.

In trackerless algorithms, since no information about the probe's pose is available, frame alignment is based on the analysis of image content, which can be achieved using either classical or DL techniques.

The classical approach typically involves two phases: in-plane and out-of-plane motion estimation. In-plane motion is determined using registration algorithms, such as those that maximize correlation between consecutive frames. Out-of-plane motion estimation, on the other hand, relies on analyzing speckle patterns by constructing decorrelation curves and estimating the displacement between frames [33, 34].

In contrast, DL methods leverage neural network architectures, such as CNNs and recurrent neural networks (RNNs), to extract features from 2D images. These features enable the network to learn complex relationships between successive frames, ultimately allowing it to determine how they should be aligned for accurate 3D volume reconstruction.

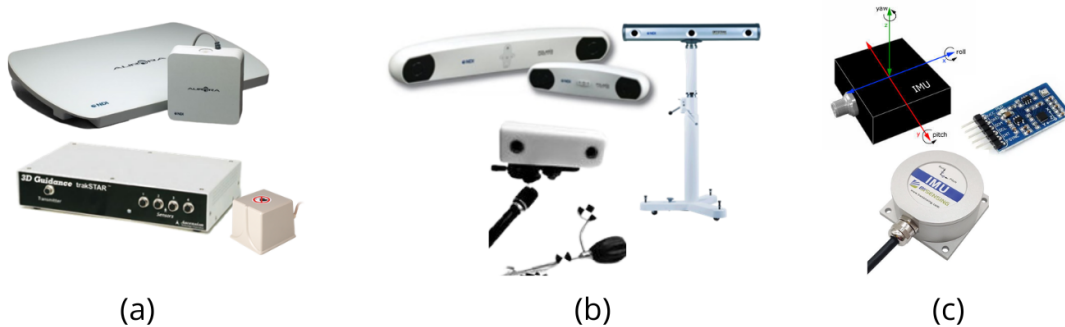


Figure 6: Available MTSs include: (a) electromagnetic trackers; (b) optical trackers; and (c) IMU sensors.

Given their numerous advantages and relevance to this dissertation, trackerless methods will receive particular focus and analysis in Section 2.3.

2.2 Volume reconstruction

After determining the position and orientation of each frame in the sequence, the next step involves interpolating the collected data to reconstruct the target volume. Interpolation methods for this process can be divided into three main categories: PBMs, VBMs, and FBMs.

2.2.1 Pixel-based methods

A commonly used PBM is the pixel nearest neighbor (PNN) method, which involves traversing each pixel in the acquired B-scans and assigning its intensity value to the nearest voxel in the 3D volume. This algorithm consists of two main stages: the bin-filling stage and the hole-filling stage [24, 35, 36].

In the bin-filling phase, pixel intensities are assigned to the nearest voxel based on the corresponding positional information. During this process, two scenarios may arise: multiple pixel intensities may be assigned to a single voxel, or conversely, some voxels may remain unmatched, with no pixel intensity assigned to them [37].

To address the first scenario, the traditional approach is to average the intensity values assigned to the specific voxel. This technique not only resolves the pixel overlap but also helps filter out speckle noise and improve the signal-to-noise ratio. Alternative methods include selecting the maximum intensity value when only a few samples are assigned to a voxel [38].

To handle the second scenario, where voxels remain unassigned, a hole-filling stage is implemented to fill these gaps using data from adjacent known voxels. Various methods have been proposed for this stage, including interpolating between the two nearest non-empty voxels, calculating their average value [36, 39], or applying an Olympic operation. The latter involves sorting the neighboring voxels, removing a certain percentage of the maximum and minimum values, and calculating the average of the remaining ones [39]. A modification proposed by Dewi *et al.* [39] introduced a thresholding step as the final stage of the Olympic operator, to account for the possibility that voxel intensities may deviate significantly from the average of their neighbors due to variations across anatomical structures.

2.2.2 Voxel-based methods

A widely used VBM is the voxel nearest neighbor (VNN) method, where the voxels in a predefined volume are traversed and filled with the intensity values from the nearest pixels based on an orthogonal projection. Unlike the PNN method, VNN accomplishes voxel filling in a single stage without requiring a subsequent hole-filling process. As an advantage, this is an intuitive and simple approach that allows the original texture patterns of the B-scans to be preserved. However, it is sensitive to misalignment between images and sparse data sampling, which can lead to discrepancies between adjacent voxels and result in significant reconstruction artifacts [21, 24, 35].

A common variant of the VNN method is the distance-weighted interpolation. This approach calculates voxel intensities by taking a weighted average of pixel intensities within a predefined ROI, with weights determined based on the distance between each pixel and the center of the voxel [24, 35, 36]. In general, this method achieves good reconstructions, with a notable reduction in speckle noise and artifacts, but often results in a more blurred volume.

Regarding how weights are assigned based on distance, various methods have been proposed, including squared distance weighting [40], median filter-based distance weighting [41], adaptive Gaussian distance weighting [24], among others.

2.2.3 Function-based methods

To improve reconstruction accuracy, several FBMs have been developed, including radial basis function (RBF) interpolation, Bézier interpolation [21, 41, 42], among others.

The RBF interpolation method attempts to estimate the underlying function that fits a set of data points [43]. For each known point, a RBF — often Gaussian — is assigned with a determined weight, representing the influence of each data point on its surrounding region in the 3D image. To estimate the continuous function, the RBFs of all points are summed as a linear combination, constructing a spline that closely aligns with the data points and ensures smooth transitions [43, 44]. Finally, once this function has been estimated, it becomes possible to interpolate the unknown values that are located between known ones. This method has demonstrated good performance by effectively suppressing speckle noise and preserving edges, resulting in higher-quality reconstructions with fewer artifacts. However, it is computationally intensive compared to the interpolation techniques mentioned earlier [44].

Bézier interpolation uses third-order Bézier curves to interpolate voxel values between four adjacent B-scans. This approach is particularly advantageous for preserving contours and maintaining high accuracy

even in the case of sparse data [42].

2.3 Trackerless volume reconstruction

This section explores the various approaches for estimating transformation parameters in the absence of MTSSs. It begins with an analysis of the earliest methods proposed for this purpose, which mainly relied on speckle pattern analysis. Following this, it examines the evolution of the DL-based methods proposed for this task, highlighting their advancements aimed at improving performance.

2.3.1 Classical approaches

To estimate the trajectory of the probe and align the obtained frame sequence, two types of motion need to be considered: in-plane and out-of-plane motions. In-plane motion is the movement that occurs within the image plane and includes translation along the axial and lateral directions, as well as rotation around the elevational axis. Out-of-plane motion involves movement perpendicular to the image plane, such as elevation translation, tilting, and yawing around the axis [33, 34]. While in-plane motion can be estimated using 2D image registration algorithms, such as maximizing correlation [34] or optical flow (a method that estimates the motion between two consecutive frames by analyzing the apparent pixel motion) [45, 46], the out-of-plane motion is more challenging to determine and relies on analyzing the speckle patterns present in each frame [47]. Speckle patterns arise as a granular effect from the interference between acoustic waves and tissue. These patterns can be mathematically modeled using Rayleigh statistics, which are applicable in the case of fully developed speckle (FDS). In FDS, the speckle pattern is stabilized and predictable due to the superposition of several scattered waves. However, the problem appears in regions where the waves maintain phase correlations, known as coherent regions, which do not exhibit FDS.

Figure 7 provides a visual summary of the speckle decorrelation approach, which is explained in greater detail in the following paragraphs.

The basis of this method lies in the analysis of the resolution cell, defined as the volume surrounding a point containing scatterers that contribute to the backscattered US signal at that location [48]. Since the cell was observed to be elongated along the elevation direction, as illustrated in Figure 8, there is some overlap between the current cell and the cell in the subsequent frame [33, 48]. Using decorrelation curves, which describe how the correlation between frames changes with the distance between them, it is possible to estimate the displacement that has occurred. To infer the tilt and yaw angles arising from

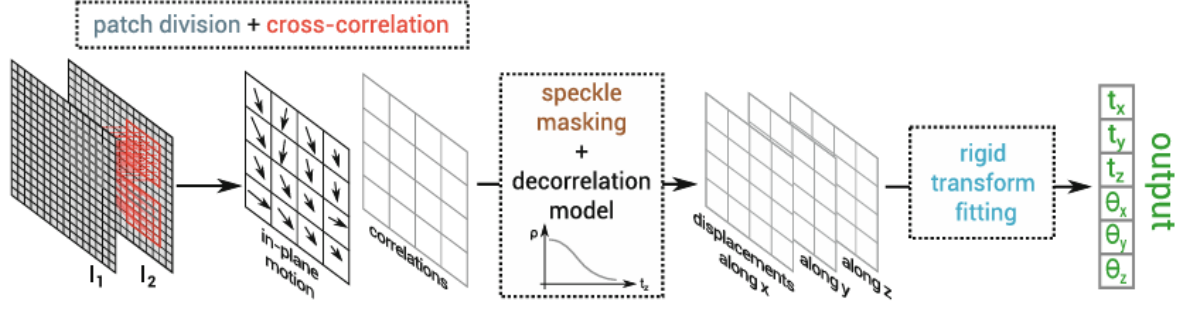


Figure 7: Overview of the speckle decorrelation framework. From [45].

the movement of the probe, the analyzed frame is divided into several patches, each associated with its own speckle decorrelation model (as the resolution cell changes with depth) [49]. Thus, the main steps of this method include determining the decorrelation curves, establishing the correlation between frames, and estimating the out-of-plane motion.

Initially, the decorrelation curve is estimated by assuming FDS patterns through mathematical derivations such as the Pearson correlation. This coefficient assumes a uniform decorrelation rate across all regions, disregarding the presence of coherent scattering regions and the variation in the resolution cell with depth, which reduces the alignment accuracy [33, 47]. Although the variation in the resolution cell can theoretically be estimated using US specifications and acoustic physics, this approach is often impractical. Instead, speckle phantoms are often used to calibrate the decorrelation curve. These phantoms are designed to replicate the acoustic properties of the biological tissues of interest, simulate their speckle patterns, and provide uniform attenuation with depth, facilitating accurate calibration of the decorrelation curve [33]. The final step in this process is to estimate the six DoFs that define the relative position and orientation between consecutive frames. This is achieved by applying a least squares fitting algorithm to the set of local 3D translation measurements obtained from all image patches (Figure 9).

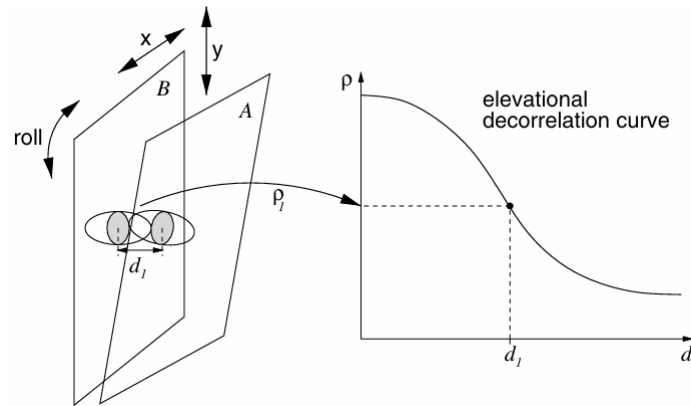


Figure 8: Fundamental concept of elevational speckle decorrelation. From [33].

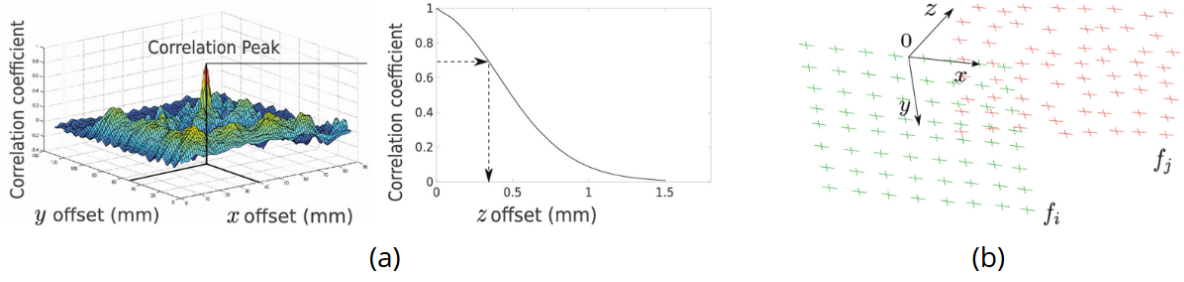


Figure 9: Illustration of the rigid motion estimation process using local 3D translation measurements. (a) Correlation analysis for translation offsets, using the correlation peak. (b) Trajectory estimation between consecutive frames using patch-based measurements. Adapted from [34].

Various strategies have been proposed to improve the previous approach and overcome the limitations observed. Chang *et al.* [50] proposed to use an additional orthogonal B-scan acquisition (Figure 10) to provide a different perspective of the scanned volume, thereby enabling error correction. First, the out-of-plane motion is estimated using a speckle decorrelation curve [33, 34]. To model this curve, a speckle detector identifies regions in the image where speckle patterns are prominent. Subsequently, an orthogonal frame is acquired relative to the freehand scanning images, resulting in the intersection of a line in each of the frames. By analyzing the location of this line, it is possible to estimate the position of each frame relative to the reference image. Finally, the positions estimated through the speckle analysis are slightly adjusted according to the locations obtained from intersection analysis. It is important to highlight that this method assumes that the acquired frames and the reference frame are perpendicular — a condition that cannot always be guaranteed.

Subsequently, Gee *et al.* [33] introduced an adaptive speckle decorrelation to refine the calibration curves by incorporating local coherent scattering. In this approach, the speckle phantom is scanned

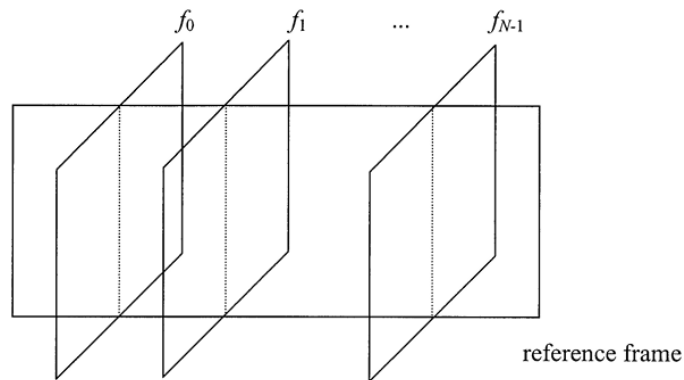


Figure 10: Illustration of the approach by Chang *et al.* [50], which includes a reference frame to improve out-of-plane motion estimation. From [50].

not only in the elevational direction but also in the axial and lateral directions. To obtain the calibration parameters, it is assumed that coherent scatterers exhibit physical isotropy, implying similar behavior across all directions. This assumption facilitates the estimation of the amount of coherent scattering using a heuristic model, resulting in improved accuracy in 3D reconstruction compared to conventional methods that consider FDS.

One limitation of the previously discussed methods is their assumption that the probe's movement is linear and monotonic, which may not be the case with freehand motion. To address this, Housden *et al.* [48] developed a novel approach capable of handling unconstrained freehand data, including nonmonotonic probe motion and frame intersections (Figure 11). The proposed algorithm involves selecting a set of widely spaced frames to reduce the number of estimations needed between them, thereby mitigating error propagation. Subsequently, the pattern of elevation displacement between two frames is analyzed to determine whether there is an intersection. For this stage, the authors incorporated the adaptive speckle decorrelation technique proposed by Gee *et al.* [33] to compensate for coherent scattering in real tissue. Finally, to verify whether the movement is nonmonotonic, the method processes three consecutive frames at a time, considering both possible directions and selecting the result that yields the smallest residual error from the least squares fit.

A year later, Laporte *et al.* [49] proposed an algorithm based on the traveling salesman problem to detect nonmonotonic motion through patch analysis. The authors suggested a further adjustment to the technique used by Housden *et al.* [48], incorporating multiple correlations using a maximum likelihood framework. This approach resulted in better estimation of out-of-plane motion and improved reconstruction accuracy.

Afsham *et al.* [47], instead of focusing solely on identifying FDS regions as done in [50], proposed a second-order Rician Inverse Gaussian (RiIG) statistical model capable of estimating out-of-plane motion for FDS and non-FDS in every patch. To achieve this, the RiIG model's statistical patch parameters were determined using an expectation-maximization algorithm. Based on these parameters, a correlation coefficient of the speckle pattern was derived from its statistical properties. In addition, the authors introduced a lin-

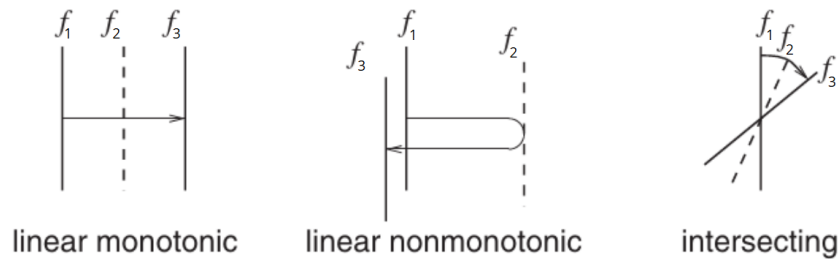


Figure 11: Example movement types encountered during a freehand scan. Adapted from [48].

ear regression model to account for changes in the beam profile, such as different widths of the resolution cells. The results demonstrated a smaller error in out-of-plane motion estimation, consistently surpassing the performance of the method proposed in [33].

Tetrel *et al.* [34] proposed a method to reduce error accumulation caused by systematic estimations of frame positions relative to the first frame. This approach, however, assumed FDS and monotonic probe motion. After estimating in-plane motion by maximizing correlation and out-of-plane motion by analyzing speckle decorrelation, a graph-based representation was introduced to optimize probe trajectory, as illustrated in Figure 12. The latter step begins by generating multiple random trajectories across the frame sequence. Starting with specified initial and final frames, an intermediate frame is randomly selected for each trajectory. The shortest path connecting the first and final frames through the intermediate frame is identified, and this process is repeated multiple times to produce various trajectory estimates. These trajectories are then averaged using an optimization-based method, with the trajectory having the overall shortest path serving as the initial estimate. This approach demonstrated improved trajectory estimation accuracy, resulting in qualitatively better volumetric reconstruction.

2.3.2 Deep learning approaches

Recent progress in DL has transformed the field of medical imaging through robust, feature-rich, and efficient solutions. Unlike classical methods, which rely on physical models and predefined assumptions like speckle pattern analysis, DL approaches learn spatial and intensity patterns directly from the input data. The analysis of richer features allows for a better assessment of frame alignment parameters, ultimately improving 3D volume reconstruction.

Prevost *et al.* [45] were the first to propose using a CNN (Figure 13) to analyze image features such as speckle patterns and determine the rigid transformations between a pair of images, improving pose

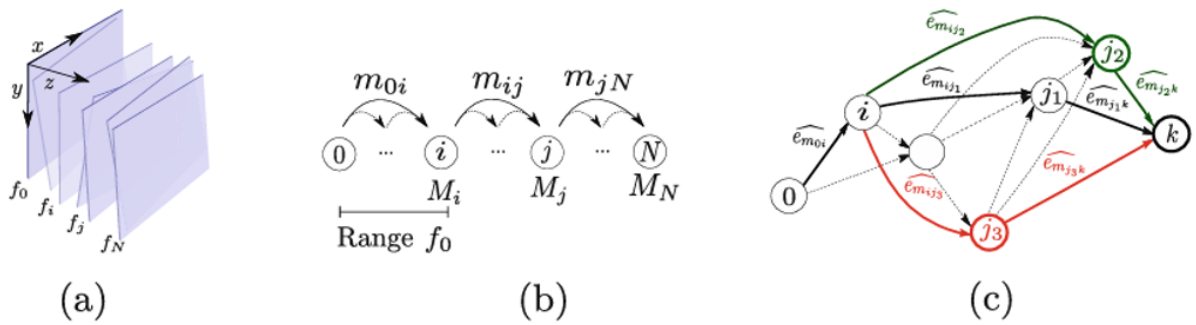


Figure 12: Freehand volume reconstruction framework proposed by Tetrel *et al.* [34]: (a) data acquisition; (b) trackerless motion estimation; and (c) trajectory optimization using graph-based methods. From [34].

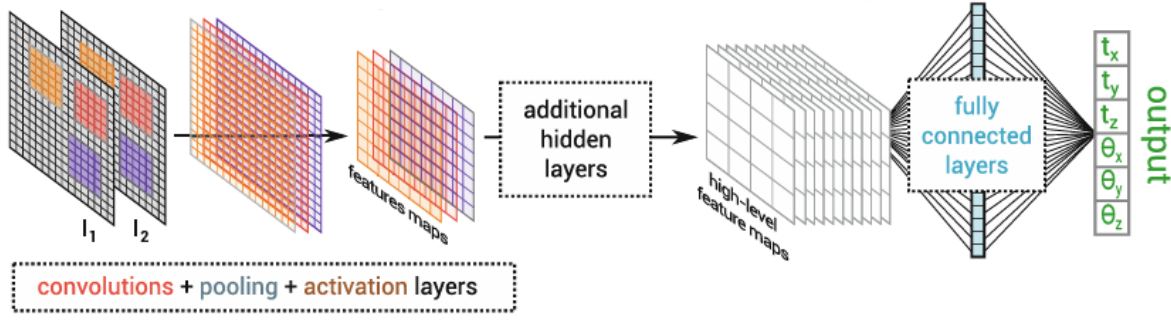


Figure 13: General workflow of CNN-based methods for pose estimation. From [45].

estimation accuracy.

CNNs have become widely used in image analysis, demonstrating effectiveness in addressing complex challenges in this field. These networks are composed of four distinct types of layers: convolutional, pooling, activation, and fully connected (FC) layers. Convolutional layers are designed to process the input data, extract spatial features, and consequently detect patterns. Pooling layers reduce dimensionality, decreasing computational load, while activation layers apply a nonlinear function to introduce nonlinearity to the model. Finally, the FC layers aggregate outputs from the previous layers and produce the final result, which in this case corresponds to the six translation and rotation parameters that describe the probe's movement between successive frames [51].

As with classical approaches, the authors emphasize that the main challenge lies in estimating out-of-plane motion. To address this, they proposed a network with four input channels: two of them being successive frames, and the other two being in-plane motion vectors estimated using optical flow. This approach allows the network to focus on out-of-plane motion, leading to notable performance improvements. The results showed significant reductions in error drift, outperforming classical methods. The authors also evaluated the impact of speckle noise on the CNN's performance and its ability to generalize. Testing the network on both filtered and unfiltered images revealed that the CNN performs better with speckle patterns present, although results with filtered images still surpassed those of the classical methods. Regarding generalization to other anatomies, it was found that the method estimates in-plane motion accurately, but shows drastically inferior performance for out-of-plane motion. In a subsequent study [29], Prevost *et al.* enhanced their approach by integrating orientation data from an IMU as input to the CNN, further improving the motion estimation accuracy.

Miura *et al.* [52] noted possible accuracy issues in the estimation of in-plane motion using the Farneback optical flow implemented in [45]. To address this, they proposed to estimate the optical flow using a pre-trained DL-based architecture, FlowNetS, incorporating the extracted motion features into their CNN model.

Additionally, a ResNet34 architecture was implemented to extract texture features. The ResNet is a type of CNN that aims to mitigate the problem of vanishing gradients – where gradients diminish as they are backpropagated through increasingly deep networks – by improving feature extraction through the use of residual connections that allow information and gradients to bypass intermediate layers [53]. As illustrated in Figure 14, the outputs of the FlowNetS and ResNet34 modules are then reduced using global average pooling (GAP) layers and concatenated before being passed through the final two FC layers. The authors also introduced a novel consistency loss function based on reprojection consistency, in which the rotation matrix and translation vector are calculated for both forward and reverse motions between consecutive frames. This method verifies whether a given point returns to its original position, thereby enhancing the precision of motion estimation between US image frames and outperforming the method introduced in [45].

In 2020, Wein *et al.* [54] proposed to improve the method developed by Prevost *et al.* [45], aiming to reduce the out-of-plane motion estimation drift between frames. To achieve this, two perpendicular freehand US acquisitions, typically transverse and sagittal sweeps, are combined to obtain information about the orthogonal distance to the imaging plane (Figure 15). This is possible because the out-of-plane distances in one sweep correspond to the in-plane distances in the other sweep. The proposed method begins by estimating the probe trajectory for each individual sweep using the CNN developed by Prevost *et al.* [45]. Subsequently, a linear co-registration is performed, followed by trajectory refinement using image-based optimization. To match the overlapping regions between perpendicular scans, an initial alignment is conducted using segmented volumes. For evaluation, the ground truth (GT) segmentation was compared

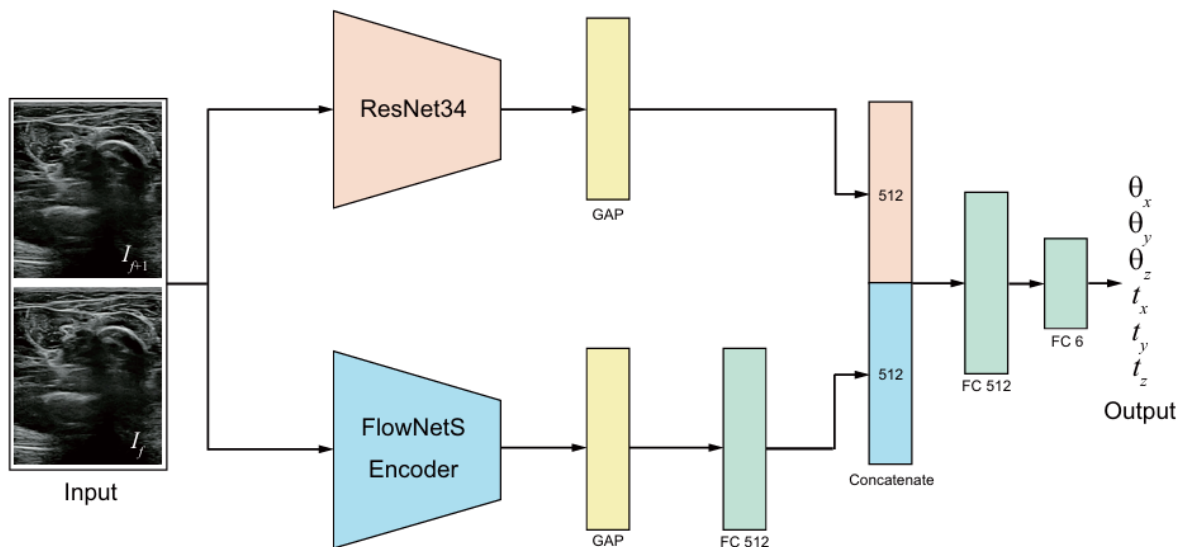


Figure 14: CNN architecture proposed by Miura *et al.* [52]. From [52].

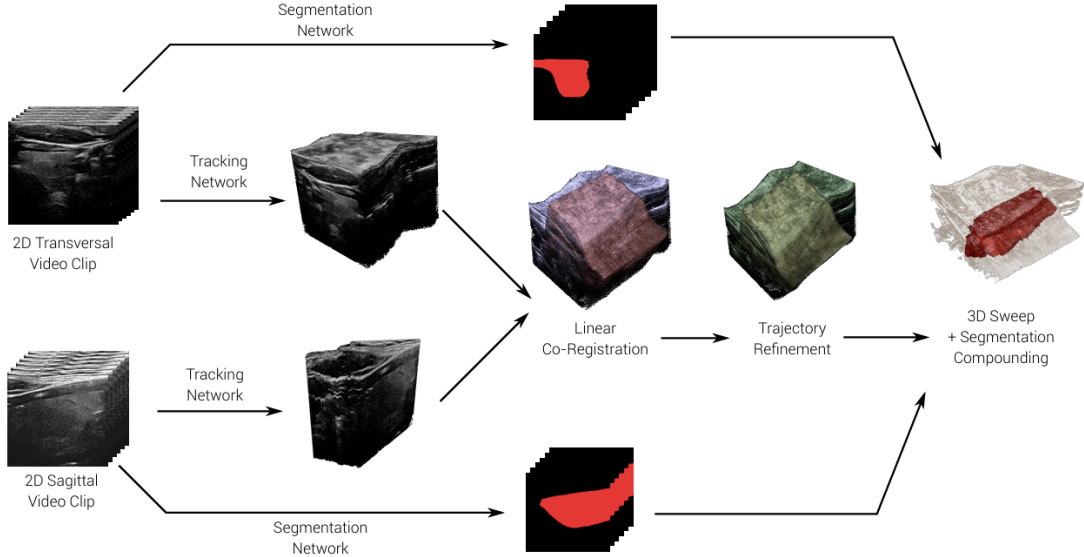


Figure 15: Workflow of the approach proposed by Wein *et al.* [54], that uses orthogonal freehand clips to improve trajectory estimation. From [54].

with the segmentation derived from the reconstructed volume. The results demonstrated a reduction in the median error by approximately 30% compared to other deep learning methods.

Prior approaches suffer from a significant limitation: they estimate transformation parameters independently for each frame pair, thereby losing the temporal information available in the recorded data. To overcome this, Guo *et al.* [55] proposed a deep contextual learning network (DCL-Net) (Figure 16), which processes an entire US video as input. By leveraging richer contextual information, this method significantly improves reconstruction accuracy. To further enhance image alignment, the authors integrated an analysis of speckle-rich regions, as done in classical approaches. This was achieved through a self-attention module, which processes feature maps from the residual block and generates an attention map that prioritizes informative regions by assigning them higher weights. Additionally, a novel case-wise correlation loss, combining mean squared error (MSE) and Pearson correlation, was introduced to prevent overfitting. The proposal suggests using both functions since the isolated use of MSE causes motion smoothing and leads to memorization of the probe trajectory. By incorporating the Pearson correlation, the unique motion pattern of a scan is emphasized, addressing the aforementioned problem. Their method achieved superior performance in terms of distance error and final drift metrics compared to the approaches in [29, 50].

Two years later, aiming to enhance probe trajectory estimation and based on their earlier work [55, 56], Guo *et al.* [57] introduced a margin ranking loss to their previous method (Figure 17). This enhancement addressed several limitations associated with supervised learning, leading to low generalization performance. The proposed strategy, termed deep contextual-contrastive network, encourages the discriminative module to learn features capable of identifying US frames with similar transformation parameters,

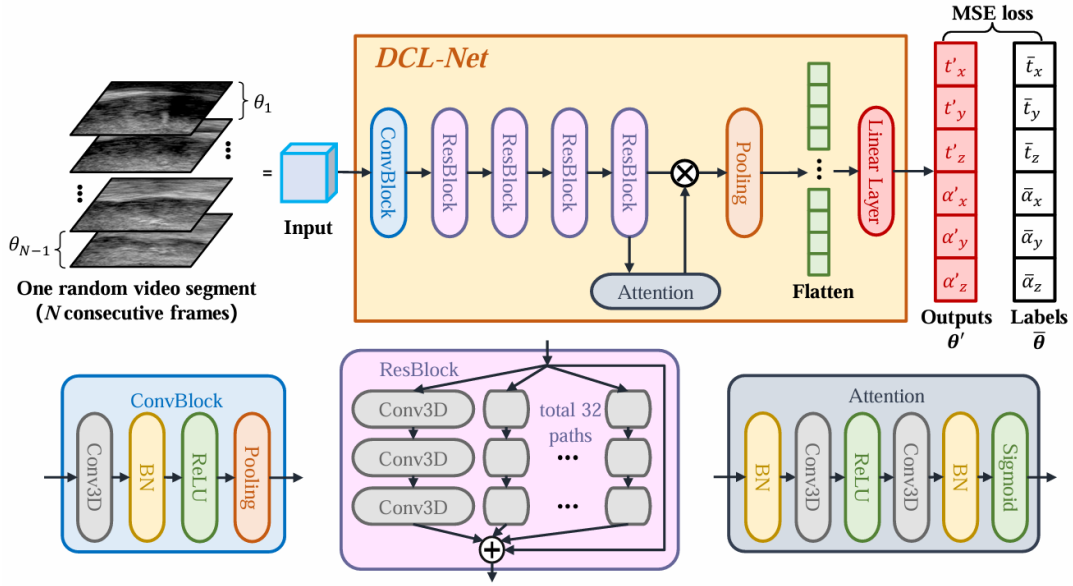


Figure 16: DCL-Net architecture proposed by Guo *et al.* [55]. From [55].

allowing the extraction of both speed and orientation from US data. For that, the margin ranking loss calculates the distance between the feature vectors of the anchor and a near sample, as well as between the anchor and a far sample, ensuring that the anchor's feature vector remains closer to the near sample than to the far sample, promoting accurate feature learning. This framework achieved superior performance compared to existing methods, including DCL-Net [55], resolving the limitation of smoothed trajectories observed in the previous approach.

Similar to Guo *et al.* [55], Li *et al.* [58] argue that using only a pair of frames to feed the network is not the most effective approach. To address this limitation, the authors proposed a multi-task learning paradigm that predicts transformations between past, current, and future frames (Figure 18). To this

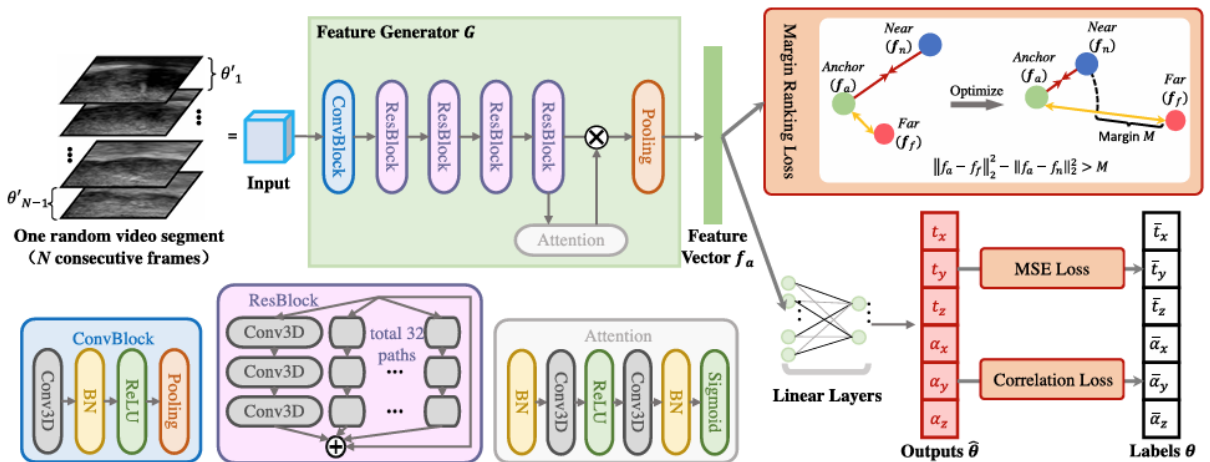


Figure 17: Modified contrastive-based DCL-Net architecture proposed by Guo *et al.* [57]. From [57].

end, the use of a CNN or a LSTM (a variant of RNNs) was proposed as the network architecture, with a comparison between the two conducted to determine which yields the best results. RNNs can process sequential data by considering information from prior states. These networks consist of a series of non-linear units, where at least one connection between units forms a directed cycle, providing continuity in predictions [59]. However, initial RNN variants faced a major limitation: a phenomenon known as the vanishing gradient problem [60, 61]. To overcome this, the LSTM was introduced, as it allows the network to retain relevant information while forgetting unnecessary data [61]. In terms of performance, the network's output estimates the transformation parameters for the original frames, as well as the auxiliary frames, which enhances generalization, improves prediction accuracy, and ensures temporal consistency of the model. During evaluation, the network exhibited lower errors when provided with a short sequence of frames, ranging from 3 to 9 frames, suggesting that overly long intervals may adversely affect temporal consistency. Furthermore, the use of LSTM produced better results compared to CNNs, confirming the effectiveness of RNNs in capturing temporal features. Overall, the authors successfully achieved their objective, highlighting the importance of temporal information in predicting spatial transformations.

Occasionally, holes are observed in 3D volume reconstructions, often caused by artifacts or fast scanning during US acquisition. To tackle this issue and enable efficient US reconstruction, Dong *et al.* [32] proposed a 3D partial convolution network (Figure 19). The method includes a 3D partial convolutional layer that uses a mask, generated using the Delaunay algorithm, to identify the location of holes. This mask is updated during each partial convolution. The authors then introduced the idea of using a generative adversarial network (GAN), an unsupervised learning approach, to fill the missing regions by learning

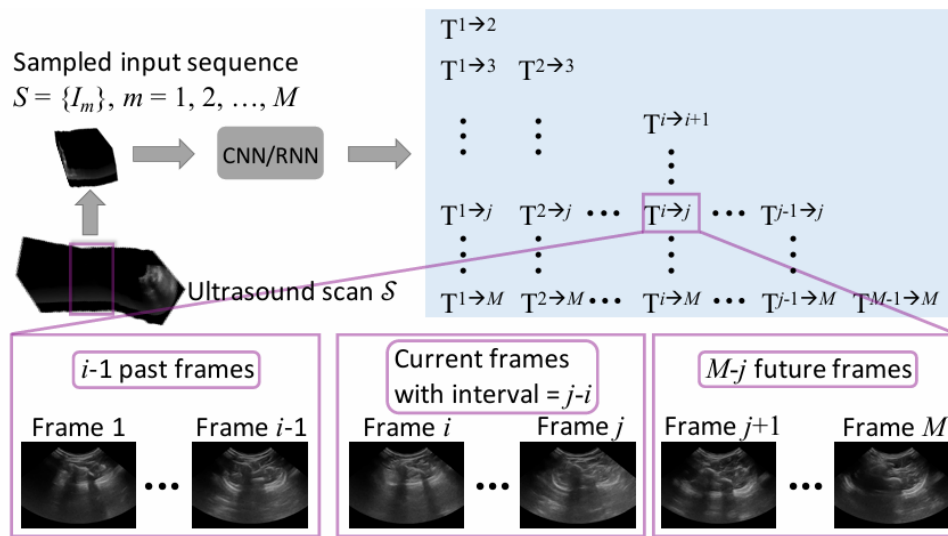


Figure 18: Multi-task learning approach proposed by Li *et al.* [58] to predict transformations for principal and auxiliary frames.

From [58].

the distribution of known regions. A GAN consists of two sub-models: a generator, which creates realistic samples to fool the discriminator, and a discriminator, which distinguishes between generated and real inputs, thereby refining the model's performance [62, 63]. To improve stability, training speed, and the quality of reconstructed volumes, the authors proposed a spectral-normalized least squares GAN. The loss function combined three terms: context loss, total variation loss, and feature-matching loss. The context loss ensures that voxels near unknown regions are weighted more heavily than those farther away. The total variation loss promotes image smoothness, while the feature-matching loss ensures structural similarity across frames. Results suggest that the proposed partial convolution network is a promising approach, yielding significant improvements in 3D structural similarity index metric (SSIM) and 3D gradient magnitude similarity deviation (GMSD).

One of the most significant issues with DL techniques is their limited generalization to different image acquisition settings, often resulting in a notable decline in performance [56, 64, 65]. Zheng *et al.* [64] encountered this issue when training a CNN with synthetically generated data and subsequently evaluating it on clinical data. To address this, the authors proposed a pairwise domain adaptation module, leveraging the observation that images at the same angle—whether simulated or real—should exhibit similar patterns. Mahapatra *et al.* [65] tackled this limitation by employing GANs to process latent feature space vectors derived from pairs of frames, generating registered images. Guo *et al.* [56] also observed generalization issues when analyzing the same anatomical structure using different transducers and motion trajectories. To address this issue and building on the previous methods, Guo *et al.* introduced a novel paired-sampling strategy combined with feature discrepancy minimization. The proposed approach notably reduced the average distance error and the final drift compared to the approaches presented in [64, 65].

A limitation of the previously mentioned studies is that they are validated only for linear scan motions,

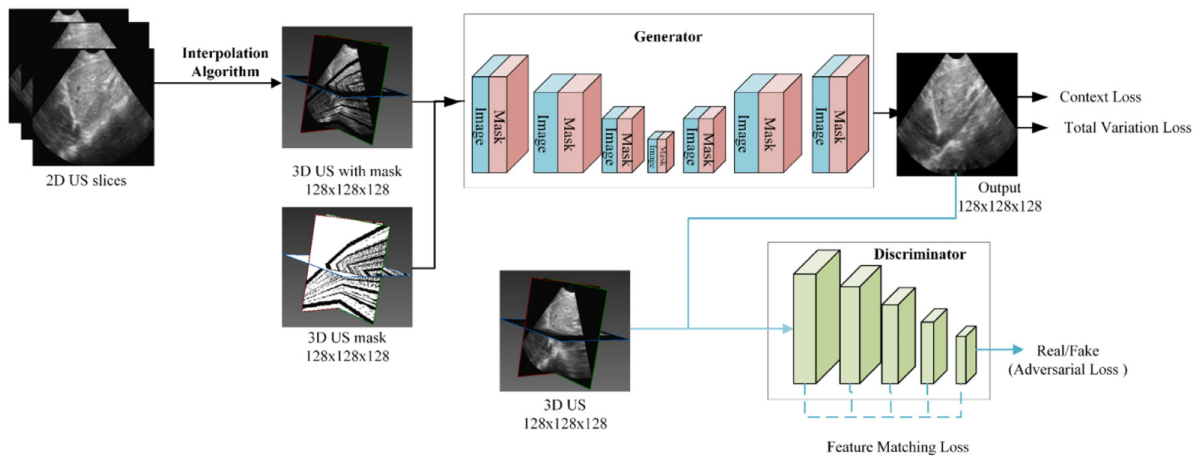


Figure 19: 3D partial convolution network proposed by Dong *et al.* [32] for freehand volume reconstruction. From [32].

while pose estimation is more challenging when the probe undergoes variations in its trajectory, such as loop scans, fast and slow scans, or sector scans. To tackle this, Luo *et al.* [66] introduced an on-line framework for freehand 3D US reconstruction that accommodates more complex scanning strategies. The proposed framework (Figure 20) integrates: self-supervised learning (SSL) and adversarial learning (ADL). The SSL component ensures that the reconstructed 3D volume preserves coherence with the original frames by extracting a slice and verifying its similarity to the original image. By incorporating this mechanism, the pose transformation parameters are predicted with awareness of the context provided by neighboring frames [66]. The ADL module is included to improve the learning of anatomical shapes, preventing irregular reconstructions. The pipeline uses a recurrent convolutional long short-term memory (ConvLSTM), a type of RNN, which enables better adaptation to the transducer's motion. For evaluation, the authors compared the final drift rate, average drift, maximum drift, and sum of drifts from their method with those reported in [29, 55], showing that their approach outperforms previous models and effectively handles complex 3D US trajectories.

Two years later, the same authors extended their approach to address diverse scanning velocities [31]. To do it, a motion-weighted training loss was proposed, combining three loss terms: mean absolute error (MAE), Pearson correlation, and motion loss. MAE compares the difference between the estimated motion

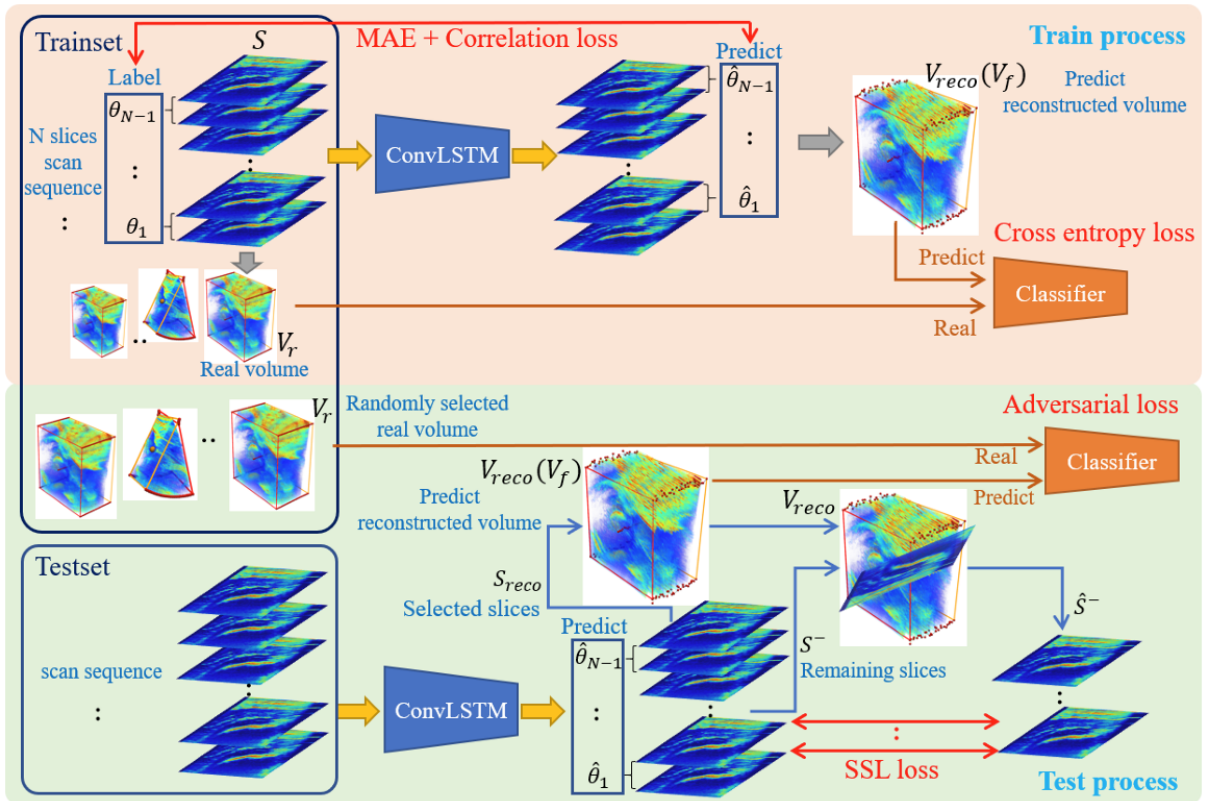


Figure 20: Online learning framework proposed by Luo *et al.* [66]. From [66].

parameters and the GT. However, alone it does not account for temporal information, leading to increased drift accumulation. Pearson correlation allows global correlation, minimizing cumulative deviation, while the motion loss term enhances the network's adaptability to variations in scan speed. The employed network involves a ResNet-LSTM cascade backbone to learn both temporal and spatial features from the 2D US frames. It basically consists of three modules: frame-level contextual consistency (FCC), path-level similarity constraint (PSC), and global adversarial shape prior (GAS) [31]. The FCC module allows for improved estimation of transformation parameters by using the reconstructed volume to analyze contextual cues. In the PSC module, the previous known scan paths are leveraged to improve network performance by utilizing similar local contextual information. Finally, the GAS module integrates a pre-trained discriminator to differentiate reconstructed volumes from real ones during training, ensuring the generation of volumes indistinguishable from actual data [31]. The pipeline of the proposed model, termed RecON, is illustrated in Figure 21. RecON outperformed the aforementioned state-of-the-art methods [29, 55, 66], emphasizing the importance of contextual cues and online learning.

With the goal of developing a stretched reconstruction of the femoral artery from 2D US images, Leblanc *et al.* [67] proposed a method that begins with segmenting the artery to align its barycenters, thereby estimating the in-plane motion. The aligned frames are then fed into a CNN to estimate the out-of-plane motion, allowing 3D volume reconstruction (Figure 22). The proposed method obtained good results in terms of final drift and average reconstruction error. However, it necessitates a pre-trained network capable of accurately segmenting the arteries. To assess the impact of this requirement, the

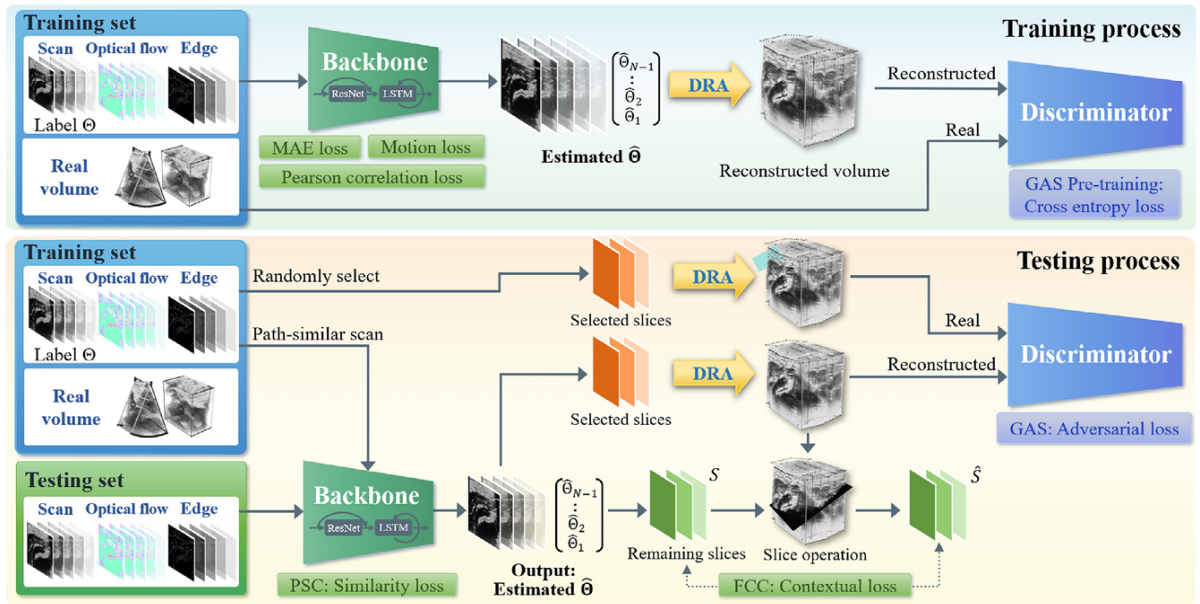


Figure 21: Improved online learning framework proposed by Luo *et al.* [31], able to handle variable scan speeds. From [31].

authors also tested estimating the out-of-plane motion with the CNN without pre-alignment for in-plane motion, observing a similar average absolute error but a noticeably higher drift.

Dou *et al.* [68] advocate that existing methods still struggle to effectively estimate out-of-plane motion, leading to reduced accuracy in volume reconstruction. To address this issue, they proposed a physics-guided learning-based approach to predict pose information for volume construction. The model incorporates physics-inspired principles, namely a correlation operator for temporal modeling via speckle decorrelation analysis, alongside 2D convolutions to capture frame-level features. This is based on the observation that the correlation decreases as the temporal distance between frames increases due to decreased cell overlap. As depicted in Figure 23, the method employs a two-stream network to separately process spatial and temporal features, which are then fused and fed into a FC layer. Using both phantom and human subject data for evaluation, the study demonstrated superior performance and computational efficiency compared to the methods presented in [55, 57, 66]. The authors further proposed sampling input frames during training in such a way that they capture longer-range temporal information, which enhances the network’s overall performance.

All previously cited studies assume that motion between each frame pair of the sequence is rigid, estimating only the six transformation parameters corresponding to the six DoFs. However, during US scanning, tissue deformation may occur due to the transducer’s pressure or involuntary tissue movements. Addressing this limitation, Li *et al.* [30] proposed a co-optimization algorithm to estimate both rigid and nonrigid transformations between frames (Figure 24). The process begins by inputting the 2D US sequence into a convolutional network to predict rigid transformations relative to a reference frame, supervised by GT data obtained via a MTS. Subsequently, an interpolation step reconstructs the volume,

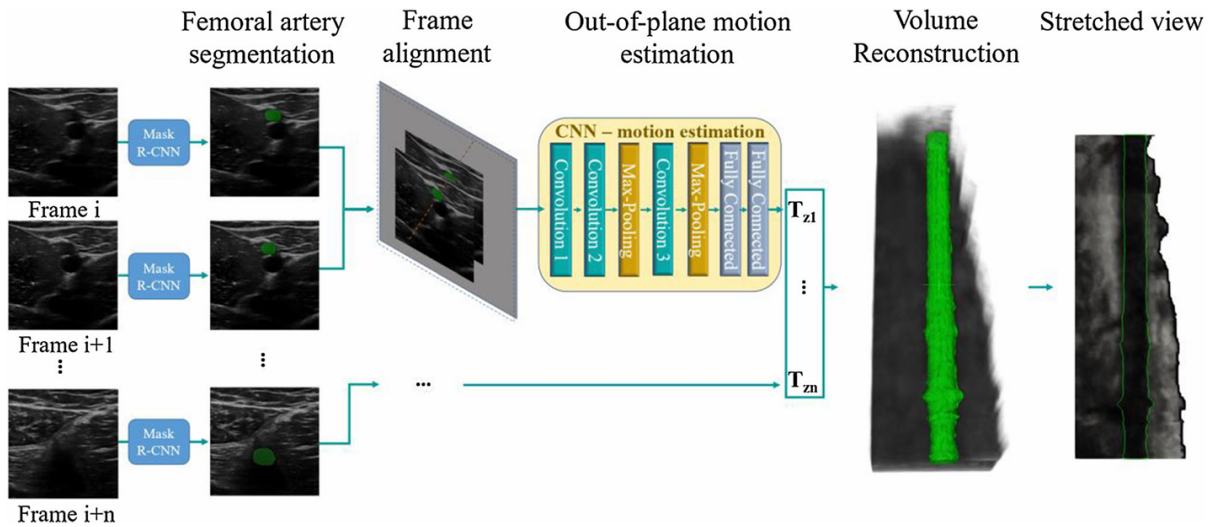


Figure 22: Workflow proposed by Leblanc *et al.* [67]. From [67].

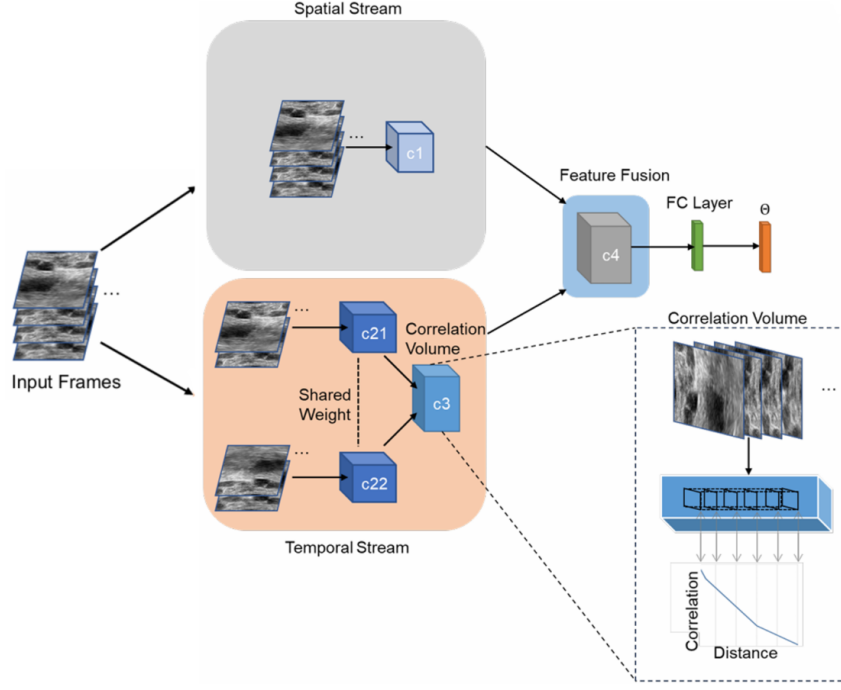


Figure 23: Overview of the physics-guided, learning-based method for pose prediction proposed by Dou *et al.* [68]. From [68].

which is then processed by the network to generate a dense displacement field. This field refines the rigidly reconstructed volume to account for nonrigid deformations. The study also explores the best training strategy to enhance network generalization. Initially, a meta-learning strategy was tested, where the rigid and nonrigid transformation stages were optimized separately. However, this approach yielded suboptimal results. Instead, the authors employed a weighted loss function to co-optimize both networks, dynamically adjusted based on the magnitude of their gradients. Using a large dataset of around 357,000 frames from 60 patients, the method demonstrated reduced global reconstruction errors, improved generalization capabilities, and introduced a robust approach for compensating nonrigid motion.

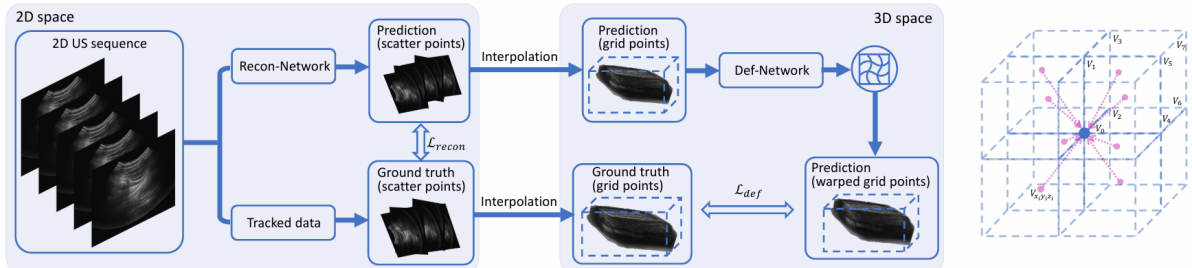


Figure 24: Workflow proposed by Li *et al.* [30] for the rigid and nonrigid motion estimation in freehand US. From [30].

2.4 Evaluation metrics

This section outlines the key metrics used to evaluate trajectory and volume reconstruction errors by comparing predicted outcomes with the GT. These metrics provide a comprehensive assessment of a model's performance in different contexts and facilitate comparisons between the results achieved by different methodologies.

2.4.1 Frame-to-frame error

One of the most important evaluation metrics is the accuracy of the estimated translation and rotation transformation parameters between each pair of frames. A commonly used measure to assess it is the MAE [52], defined in Equation 2.1. This metric is calculated by averaging the absolute difference between the estimated transformation parameters $(T_k(i \rightarrow j)^{\text{pred}})$ and the corresponding GT parameters $(T_k(i \rightarrow j)^{\text{GT}})$ for each frame pair k across a sweep (composed of $N - 1$ pairs), where i and j represent the successive frames under analysis, and N is the total number of frames.

$$\text{MAE}(t_x, t_y, t_z, \theta_x, \theta_y, \theta_z) = \frac{1}{N-1} \sum_{k=1}^{N-1} |(T_k(i \rightarrow j)^{\text{pred}} - T_k(i \rightarrow j)^{\text{GT}})| \quad (2.1)$$

Similarly, the root mean squared error (RMSE) [68] can also be employed, being expressed as:

$$\text{RMSE}(t_x, t_y, t_z, \theta_x, \theta_y, \theta_z) = \sqrt{\frac{1}{N-1} \sum_{k=1}^{N-1} (T_k(i \rightarrow j)^{\text{pred}} - T_k(i \rightarrow j)^{\text{GT}})^2} \quad (2.2)$$

Another widely used evaluation metric is the distance error (Equation 2.3) [55–57]. This metric computes the average Euclidean distance between the four corner points of the ROI defined in the input image, providing an assessment of the positional accuracy of the frame sequence, rather than focusing solely on individual transformation parameters.

$$\text{Distance Error} = \frac{1}{N} \sum_{i=1}^N \left(\sum_{c=1}^4 \sqrt{(x_{i,c}^{\text{pred}} - x_{i,c}^{\text{GT}})^2 + (y_{i,c}^{\text{pred}} - y_{i,c}^{\text{GT}})^2 + (z_{i,c}^{\text{pred}} - z_{i,c}^{\text{GT}})^2} \right) \quad (2.3)$$

where $x_{i,c}^{\text{pred}}$, $y_{i,c}^{\text{pred}}$ and $z_{i,c}^{\text{pred}}$ correspond to the coordinates of a determined point in the estimated frame, and $x_{i,c}^{\text{GT}}$, $y_{i,c}^{\text{GT}}$ and $z_{i,c}^{\text{GT}}$ represent its real position in the GT, with c being the index of the four corners, and i is the index of the frame under analysis.

Cumulative trajectory estimation error

To evaluate the cumulative estimation error, drift metrics are commonly employed, including final drift (FD), final drift rate (FDR), maximum/minimum drift, and average/median drift. In all cases, a lower drift is indicative of a more accurate reconstruction.

The FD (Equation 4.9) measures the distance between the last frame's position and its expected position based on the GT, capturing the accumulated error over the entire sweep [31, 67, 68]. Additionally, FDR (Equation 2.5) assesses the relative error by normalizing FD by the GT length of the sweep, with the result expressed as a percentage. The remaining four drift metrics follow a similar concept: maximum and minimum drift (Equations 2.6 and 2.7) refer, respectively, to the largest and smallest distances between the estimated frame positions and their respective GT positions, while mean and median drift (Equations 2.8 and 2.9) describe, respectively, the average and median errors over all frames in a sweep.

$$\text{FD} = \|\vec{t}_N - \vec{t}_N^{\text{GT}}\| \quad (2.4)$$

$$\text{FDR}(\%) = \frac{\text{FD}}{\|\vec{t}_N^{\text{GT}} - \vec{t}_1^{\text{GT}}\|} \times 100 \quad (2.5)$$

$$\text{maximum drift} = \max_n \|\vec{t}_n - \vec{t}_n^{\text{GT}}\| \quad (2.6)$$

$$\text{minimum drift} = \min_n \|\vec{t}_n - \vec{t}_n^{\text{GT}}\| \quad (2.7)$$

$$\text{mean drift} = \frac{1}{N} \sum_{n=1}^N \|\vec{t}_n - \vec{t}_n^{\text{GT}}\| \quad (2.8)$$

$$\text{median drift} = \text{median} (\|\vec{t}_n - \vec{t}_n^{\text{GT}}\|) \quad (2.9)$$

where $\vec{t}_N$ denotes the estimated final position of the last frame in the sequence, and $\vec{t}_N^{\text{GT}}$ refers to the GT position of that same frame, $\vec{t}_n$ and $\vec{t}_n^{\text{GT}}$ represent the estimated and true positions for the frame indexed by n , respectively, while $\vec{t}_1^{\text{GT}}$ is the GT position for the first frame of the sweep. Both the estimated and the reference position of the frame are with respect to the frame's origin.

In contrast, the Hausdorff distance (HD) (Equation 2.10) measures the greatest distance between any point a on the estimated trajectory and the closest point q on the actual trajectory (and vice-versa), highlighting the maximum deviation across all frames in the sequence [31, 66].

$$\text{HD}(\vec{t}_{tr}, \vec{t}_{tr}^{\text{GT}}) = \max \left\{ \max_{a \in \vec{t}_{tr}} \min_{q \in \vec{t}_{tr}^{\text{GT}}} d(a, q), \max_{q \in \vec{t}_{tr}^{\text{GT}}} \min_{a \in \vec{t}_{tr}} d(a, q) \right\} \quad (2.10)$$

where $\vec{t}_{tr}$ represents the estimated trajectory, and $\vec{t}_{tr}^{\text{GT}}$ the real trajectory.

However, since the HD is highly sensitive to outliers, a commonly used alternative is the 95th percentile Hausdorff distance (HD95). Rather than considering the maximum of all point-to-set distances, the HD95 computes the 95th percentile of the distance distribution, providing a more robust measure of the typical worst case deviation while reducing the influence of isolated large errors.

Two additional indicators have been proposed to evaluate the cumulative error: the relative trajectory error and the relative length error, shown in Equations 2.11 and 2.12, respectively. The first metric measures the cumulative in-plane translation error at each frame of the sweep, normalized by the GT sweep length. The second metric compares the estimated scan length with the GT length, normalizing it by the GT length, and assesses the method's accuracy in determining the overall sweep length.

$$\text{relative trajectory error} = \frac{\sum_{n=1}^N \|\vec{t}_n - \vec{t}_n^{\text{GT}}\|}{\|\vec{t}_N^{\text{GT}} - \vec{t}_1^{\text{GT}}\|} \quad (2.11)$$

$$\text{relative length error} = \frac{\left| \|\vec{t}_N - \vec{t}_1\| - \|\vec{t}_N^{\text{GT}} - \vec{t}_1^{\text{GT}}\| \right|}{\|\vec{t}_N^{\text{GT}} - \vec{t}_1^{\text{GT}}\|} \quad (2.12)$$

2.4.2 Volume error

In terms of volume reconstruction errors, the metrics can be divided into two categories: the analysis of reconstruction intensities and the analysis of segmented volume reconstructions. In the first approach, the intensities and gradients of the voxels in the reconstructed volume (V^{pred}) are compared to those in the GT volumetric images (V^{GT}). The second method involves segmenting the target object after volume reconstruction (s^{pred}) and then comparing it with the segmentation of the GT volume (s^{GT}). It is important to emphasize that the GT volumes are generated through a volume reconstruction process using interpolation based on spatial information obtained from a MTS.

Evaluation of volume intensity consistency

To assess structural similarity between the reconstructed and GT volumes, the SSIM can be used to evaluate the perceptual quality. This metric tries to imitate human visual perception by examining luminance, contrast, and structure between the two volumes, analyzing several patches p individually and then performing a global analysis [69]. As shown in Equation 2.13, the mathematical formulation

consists of three terms corresponding to these components. The first factor, luminance, compares the mean luminance of the predicted and reference volumes ($\mu_{V^{\text{pred}}}$ and $\mu_{V^{\text{GT}}}$, respectively). The second term, contrast, evaluates the standard deviation between the intensities of the two volumes ($\sigma_{V^{\text{pred}}}$ and $\sigma_{V^{\text{GT}}}$). The third component measures the correlation between the intensities volumes by examining their covariance ($\sigma_{V^{\text{pred}}V^{\text{GT}}}$). When the respective characteristics of the compared volumes are equal, all three factors reach a maximum value of one [70, 71]. The positive constants C_1 , C_2 , and C_3 are included to prevent a zero denominator.

$$\text{SSIM}(V^{\text{pred}}, V^{\text{GT}}) = \prod_{p=1}^M \left[\frac{(2\mu_{V^{\text{pred}}}(p)\mu_{V^{\text{GT}}}(p) + C_1)}{(\mu_{V^{\text{pred}}}(p)^2 + \mu_{V^{\text{GT}}}(p)^2 + C_1)} \cdot \frac{(2\sigma_{V^{\text{pred}}}(p)\sigma_{V^{\text{GT}}}(p) + C_2)}{(\sigma_{V^{\text{pred}}}(p)^2 + \sigma_{V^{\text{GT}}}(p)^2 + C_2)} \cdot \frac{\sigma_{V^{\text{pred}}V^{\text{GT}}}(p) + C_3}{\sigma_{V^{\text{pred}}}(p)\sigma_{V^{\text{GT}}}(p) + C_3} \right] \quad (2.13)$$

where M represents the number of patches within the volume.

Another commonly used metric is the GMSD [32], defined in Equation 2.14. This metric assesses local gradient similarities between reconstructed and GT volumes. The process begins by applying convolutional filters along each dimension (X, Y, and Z, given by h_x , h_y , and h_z) and calculating the total gradient magnitude (Equations 2.16 and 2.17). GMSD is subsequently determined as the standard deviation of gradient magnitude similarity (GMS) values (Equation 2.15) across the whole volume. As before, a constant C is included to prevent division by zero. A higher GMSD indicates greater perceptual inconsistencies and distortion in the reconstruction, while a lower GMSD reflects a better structural correspondence between reconstructed and GT volumes [72].

$$\text{GMSD} = \sqrt{\frac{1}{T} \sum_{v=1}^T \left(\text{GMS}(v) - \frac{1}{T} \sum_{v=1}^T \text{GMS}(v) \right)^2} \quad \text{where:} \quad (2.14)$$

$$\text{GMS}(v) = \frac{2m_{GT}(v) \cdot m_{pred}(v) + C}{m_{GT}^2(v) + m_{pred}^2(v) + C} \quad (2.15)$$

$$m_{pred} = \sqrt{(V^{\text{pred}} \otimes h_x)^2 + (V^{\text{pred}} \otimes h_y)^2 + (V^{\text{pred}} \otimes h_z)^2} \quad (2.16)$$

$$m_{GT} = \sqrt{(V^{\text{GT}} \otimes h_x)^2 + (V^{\text{GT}} \otimes h_y)^2 + (V^{\text{GT}} \otimes h_z)^2} \quad (2.17)$$

where T denotes the total number of voxels.

One of the issues discussed in Section 2.3.2 is the presence of holes in reconstructed 3D volumes. To address this, Dong *et al.* [32] proposed a method to accurately fill in the intensity values of unknown voxels

and introduced a corresponding evaluation metric, termed smooth neighborhood descriptor (SND). As defined in Equation 2.18, this metric assesses the smoothness of the reconstructed volume by calculating the difference between filled voxels in the edges of the reconstructed missing region (unknown voxels that have been filled) and their neighborhood (which may include both known and unknown voxels). For each filled voxel x_v , the average absolute intensity difference D_R is calculated relative to its neighbor voxels (N_v), defined within a specified search region R (six-neighborhood in Dong *et al.* [32]). The SND metric is then calculated as the average over the total number of voxels m located at the boundaries of the reconstructed missing regions. A smaller SND value indicates a greater similarity between the filled regions and its surroundings, suggesting a more gradual transition and, therefore, an improved reconstruction.

$$\text{SND}(I, x, R) = \frac{\sum_{v=1}^m D_R(I, x_v)}{m} \quad \text{where} \quad D_R(I, x_v) = \frac{\sum_{r \in N_v} |I(x_v) - I(x_r)|}{|N_v|} \quad (2.18)$$

where I denotes the voxel intensity within the US volume.

Additionally, another metric proposed to evaluate image quality was the perception-based image quality evaluator (PIQE) (Equation 2.19). This metric analyzes the image content only, starting by dividing the volume into specific regions and then evaluating the distortion in each block B_k , being capable of detecting artifacts without relying on GT. In this equation, D_{sk} represents the perceptual error for each analyzed region, and N_{SA} is the total number of active regions in the image, meaning regions that contain significant variations such as edges or significant texture [73]. It should be noted that a higher PIQE value indicates a greater number of discrepancies [32].

$$\text{PIQE} = \frac{\left(\sum_{k=1}^{N_{SA}} D_{sk}\right) + C}{(N_{SA} + C)} \quad \text{where} \quad D_{sk} = \begin{cases} 1, & \text{if } \sigma_{pq} \langle T_{STD} \& \sigma_{blk} \rangle 2 * \beta \\ \nu_{blk}, & \text{if } \sigma_{blk} > 2 * \beta \\ (1 - \nu_{blk}), & \text{if } \sigma_{pq} < T_{STD} \end{cases} \quad (2.19)$$

where C is a constant included to prevent a zero denominator, σ_{pq} denotes the structuring element at the edge of each block, T_{STD} is the threshold value, σ_{blk} is the standard deviation of each active region, ν_{blk} represents the variance of pixel intensities within a given block and β is a parameter that evaluates the connection between the center-surround deviation ratio and the standard deviation of the region [32].

Evaluation of segmented reconstructed volume

To evaluate the performance of the segmented reconstructed volume, two major groups of metrics are commonly used. The first group comprises overlap-based metrics, which assess the similarity between

segmentations by measuring the extent of their overlap. The second group consists of boundary-based metrics, which evaluate the displacement between the boundaries of the segmentations.

Overlap-based metrics

One of the most widely used metrics for segmentation analysis is the Dice coefficient [54, 57, 58, 67], defined in Equation 2.20. This metric qualifies the overlap between the segmented volume derived from the 3D reconstruction based on 2D frames and the GT volume segmentation, directly reflecting the method's accuracy. A higher Dice coefficient indicates a higher degree of overlap, with a value of one representing perfect agreement and a value of zero indicating no overlap.

$$\text{Dice}(s^{\text{pred}}, s^{\text{GT}}) = \frac{2 \cdot |s^{\text{pred}} \cap s^{\text{GT}}|}{|s^{\text{pred}}| + |s^{\text{GT}}|} \quad (2.20)$$

Another similarity measure, related to the Dice coefficient and also based on the F-measure, is the Jaccard similarity coefficient (Equation 2.21). This metric measures overlap relative to union: it is defined as the ratio between the intersection and union of the segmented volumes, and has been shown to more effectively penalize both under- and over-segmentation compared to Dice [74, 75].

$$J(s^{\text{pred}}, s^{\text{GT}}) = \frac{|s^{\text{pred}} \cap s^{\text{GT}}|}{|s^{\text{pred}} \cup s^{\text{GT}}|} \quad (2.21)$$

In order to evaluate the proportion of voxels predicted as part of the volume that actually belong to the GT segmentation, the precision metric, equation 2.22, is commonly used. This metric is defined as the ratio of true positives (TP) to the sum of TP and false positives (FP), and ranges from 0 to 1, with 1 indicating the absence of false positives. Thus, a value close to 1 means that nearly all voxels reconstructed as the target anatomy truly correspond to the anatomy in the GT.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (2.22)$$

Conversely, the recall metric, equation 2.23, evaluates the proportion of voxels of the target anatomy that have been correctly identified. It is defined as the ratio of TP to the sum of TP and false negatives (FN), rather than FP as in the previous metric. In this way, a high recall value indicates that only a small portion of the target volume has been missed.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (2.23)$$

Boundary-based metrics

The HD metric, previously described in subsection 2.4.1, can also be applied to evaluate how closely the target boundaries align in both segmentations. Here, instead of analyzing the distance between trajec-

tory points, this metric evaluates the distance from each point on the surface mesh of one segmentation to the nearest point on the surface of the other segmentation. However, this metric is highly sensitive to outliers [74], which motivates the use of a symmetric HD. In this approach, the HD is calculated in both directions, from s_{pred} to s_{GT} and vice versa, and the maximum value is considered, allowing one to infer about the local major error between the two surfaces. This provides a robust measure of the largest local deviation between the two surfaces.

$$d(s^{\text{pred}}, s^{\text{GT}}) = \frac{1}{|s^{\text{pred}}|} \sum_{p \in s^{\text{pred}}} \min_{q \in s^{\text{GT}}} |p - q| \quad (2.24)$$

$$\text{symmetric HD}(s^{\text{pred}}, s^{\text{GT}}) = \max \left(d(s^{\text{pred}}, s^{\text{GT}}), d(s^{\text{GT}}, s^{\text{pred}}) \right) \quad (2.25)$$

Similar to the symmetric HD, the average symmetric surface distance (ASSD) measures the distance in both directions. However, instead of evaluating the maximum distance from one surface point to another, it calculates the mean distance in both directions, providing a measure of how close the segmentations are overall.

$$d(x, s^{\text{GT}}) = \min_{y \in s^{\text{GT}}} \|x - y\| \quad (2.26)$$

$$d(y, s^{\text{pred}}) = \min_{x \in s^{\text{pred}}} \|y - x\| \quad (2.27)$$

$$\text{ASSD}(s^{\text{pred}}, s^{\text{GT}}) = \frac{1}{|s^{\text{pred}}| + |s^{\text{GT}}|} \left(\sum_{x \in s^{\text{pred}}} d(x, s^{\text{GT}}) + \sum_{y \in s^{\text{GT}}} d(y, s^{\text{pred}}) \right) \quad (2.28)$$

Volume-based metrics

Another approach for evaluating volume reconstruction focuses on the volume's structural characteristics [54, 76]. Following the segmentation of the target anatomy within the V^{pred} and V^{GT} , the total volume of the segmented region is calculated in milliliters for both volumes. These segmented volumes are then compared using metrics such as the relative volume difference (RVD), equation 2.29, or statistical correlations, including Pearson correlation or Spearman correlation.

$$\text{RVD}(\%) = \frac{V(s^{\text{pred}}) - V(s^{\text{GT}})}{V(s^{\text{GT}})} \times 100 \quad (2.29)$$

where $V(s)$ denotes the physical volume computed from the segmentation s .

Chapter 3

Tracker-based pose estimation

This chapter presents the methods used for MTS-based pose estimation and ensuing volume reconstruction. The proposed setup and resulting reconstructions are then validated, and the results are analyzed and discussed to evaluate the system's performance and its applicability to the proposed scenario.

3.1 Methods

As introduced in Chapter 2, to perform volume reconstruction from a sequence of 2D frames, a MTS-based method can be implemented. To accomplish this, the methodology presented in this chapter is divided into four main sections, as illustrated in Figure 25.

The first section focuses on characterizing the US system, including a detailed description of the MTS, its components and devices, as well as the frame grabber used to acquire images from the US machine.

The second section describes the calibration procedures performed to ensure synchronization between the MTS and the US system during the acquisition of the 2D freehand scans.

The third section explains the entire data acquisition process, detailing how both image and pose data, obtained respectively through the US system and the MTS, were handled. This stage also presents the preprocessing pipeline, including the mathematical adjustments required to guarantee consistency in the transformation matrices.

Finally, the fourth section details the volume reconstruction process, performed on the preprocessed data using dedicated software tools, namely the 3D Slicer (The Slicer Community, USA, version 5.8.1).

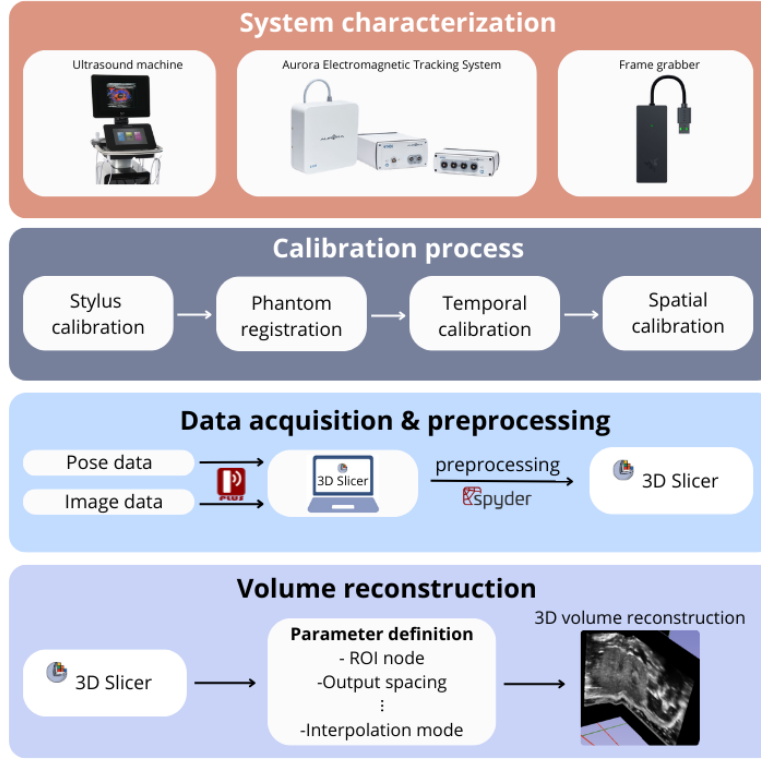


Figure 25: Overview of the proposed MTS-based framework.

3.1.1 System and hardware

US imaging: Vevo 3100

In this study, US imaging is performed using the Vevo 3100 high-resolution imaging system (Visual Sonics Inc, Canada), which is specifically designed for small animal models such as mice. The MX250 transducer is employed due to its wider field of view compared to the other available option, allowing the monitoring of tumor growth over time. The transducer operates within a frequency range of 15–30 MHz, provides an axial resolution of 75 μm , and supports a maximum imaging depth of 30 mm with a lateral field of view of 23.04 mm [77].

To acquire the US image data, a frame grabber (Razer Ripsaw X, Razer Inc., USA) is used to capture the US system's display and transmit it to the computer. This device supports video capture at up to 4K resolution at 30 frames per second (fps), which satisfies the acquisition requirements of this study.

Electromagnetic tracking: NDI Aurora

To obtain the pose data of the US probe over time, the Aurora electromagnetic MTS is employed. This technology relies on the emission of a controlled electromagnetic field with predefined characteristics, which allows the detection of the sensor's position and orientation through changes in the field. The core

components of this system are the field generator (FG), the sensor interface unit (SIU), the system control unit (SCU), and the sensors, as can be seen in Figure 26.

The components are physically connected to enable data transmission and reception, working together to detect the sensors, preprocess the signals, and transmit the pose data to the computer. The SCU controls the FG, adjusting its operating parameters and thereby defining the electromagnetic field in which the sensors are detected [78]. Each sensor is connected to the SIU, which amplifies and converts the analogue signals into digital format. The SCU then collects the digitized data from the SIU, calculates the position and orientation of each sensor, and transmits this information to the computer.

A planar 20–20 FG with a cubic measurement volume of $500 \times 500 \times 500 \text{ mm}^3$ is employed, requiring that the sensors remain within this volume for reliable detection. A 6-DoF sensor is attached to the US probe to acquire both rotational (yaw, pitch, and roll) and translational data, while another similar sensor is positioned between the fingers of the researcher’s hand used to restrain the mouse during the 2D sweep. This configuration enables the monitoring of both probe motion and any involuntary hand movements during freehand acquisition. Since the hand holding the mouse is expected to remain still throughout the procedure and only the probe should move, capturing such unintentional hand motion is essential to ensure accurate volume reconstruction. Additionally, as the sensor attached to the probe directly influences the fidelity of the recorded position and orientation, its stability during data acquisition is critical, since any displacement relative to the probe can introduce errors in the transformation matrices and compromise reconstruction quality, as discussed in this Section in the context of probe–sensor integration.

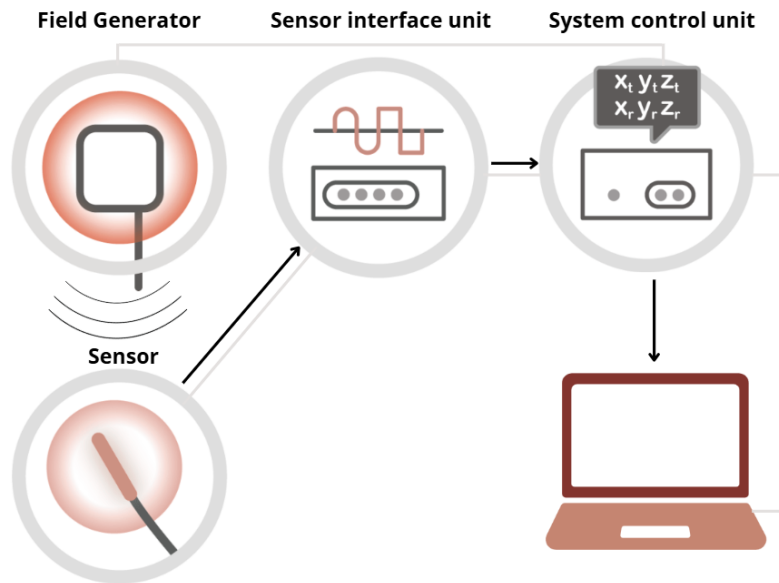


Figure 26: Data flow between the core components of the Aurora MTS. Black arrows indicate the direction of data flow, while light gray lines represent physical cable connections. Adapted from [78].

Electromagnetic interference mitigation

All acquisitions and calibration procedures were conducted at the ICVS animal facilities, in the same room where the freehand experiments were subsequently performed. The environment contains several metallic elements, including the workbench, support structures, and laboratory equipment, which could distort the electromagnetic field and reduce the accuracy of the Aurora MTS in detecting the sensor poses. To mitigate these effects, all potential interference sources were positioned as far as possible from the FG, and the tracking setup was placed on top of a large container turned upside down, preventing the Aurora FG from resting directly on the table with metallic legs. This configuration effectively minimized field perturbations and ensured that the calibration results obtained in this environment were representative of the actual acquisition conditions.

Probe-sensor integration

To ensure accurate volume reconstruction, the sensor attached to the US probe must remain stable throughout data acquisition. To achieve this, a sensor holder was designed using computer aided-design (CAD) software (SolidWorks, Dassault Systèmes SolidWorks Corporation, Waltham, Massachusetts, United States). The holder geometry was modeled according to the probe's shape and dimensions to ensure a precise fit, stable sensor positioning, and minimal interference with the operator's handling during freehand scanning. The holder was then 3D printed, with printing parameters defined using Ultimaker Cura software on an Ultimaker II Plus printer (Ultimaker B.V., Geldermalsen, The Netherlands). After fabrication, the support was rigidly attached to the probe, with the sensor remaining fixed within the structure for the entire duration of the study. Importantly, neither the sensor nor the holder was removed at any point, ensuring that their relative positioning remained consistent across all acquisitions. Appendix A presents the SolidWorks technical drawings of the probe and the holder, as well as their final appearance when mounted on the probe in the actual experimental setup.

3.1.2 Calibration procedures

Once the holder is attached to the probe and all necessary components connected, the MTS was calibrated using the Public Software Library for Ultrasound (PLUS) Toolkit. All calibrations were performed using a PLUS configuration file that specified the setup and parameters of each system component, such as ID, type, and serial port, to ensure proper identification within the PLUS environment.

These calibration steps are essential to address two main aspects: (1) the temporal synchronization

between the image data captured by the frame grabber and the pose data acquired from the sensor, as the MTS transmits data at a significantly higher rate than the frame grabber, causing the image data to arrive at the computer with a delay relative to the pose data; and (2) the spatial alignment between components, since the pose data is referenced to the coordinate system of the sensor fixed within the holder, which differs from that of the US image. Therefore, it is necessary to determine the transformation matrix that converts the image coordinate system into the sensor coordinate system.

Materials and equipment

To carry out the required calibration procedures, several components were used, as detailed below.

For the spatial calibration, the FCal-3.1 phantom was used together with a stylus equipped with an integrated 6-DoF sensor (Northern Digital, Inc. (NDI), Waterloo, Ontario, Canada) and an additional 6-DoF reference sensor (Northern Digital, Inc. (NDI), Waterloo, Ontario, Canada) attached to the phantom. The reference sensor was positioned above a printed cross marker on the phantom and remained fixed during the phantom registration and subsequent spatial calibration. To enable phantom registration, its geometry and the exact coordinates of several reference points were defined in the extensible markup language (XML) configuration file used by the PLUS Toolkit for the MTS calibrations. All distances provided refer to a predefined origin point corresponding to the virtual phantom's coordinate system. For the final spatial calibration, a set of wires arranged in a N-wire configuration were placed in the phantom, with their start and end points defined according to the phantom's numbering system. Since the US probe has a maximum depth limit of 30 mm, the wires were located only near the top surface of the phantom.

For both temporal and spatial calibration procedures [79], the phantom must be immersed in water to enable US image acquisition. Therefore, a water-filled container was used to keep the phantom fully submerged during the calibration process. To minimize US wave reflections from the container's walls, the inner surfaces were lined with cork, a material with low acoustic impedance, relatively close to that of water, which enhances wave absorption and reduces image artifacts. All components used in these procedures are shown in Figure 27.

Stylus calibration

To perform the spatial calibration, two preliminary calibrations are required. The first is the stylus calibration, which is essential for registering the phantom reference points. As mentioned earlier, the stylus already contains an integrated sensor, but it is crucial to accurately define the position of the stylus tip relative to this sensor, since the points on the phantom are identified through the tip. During calibration,

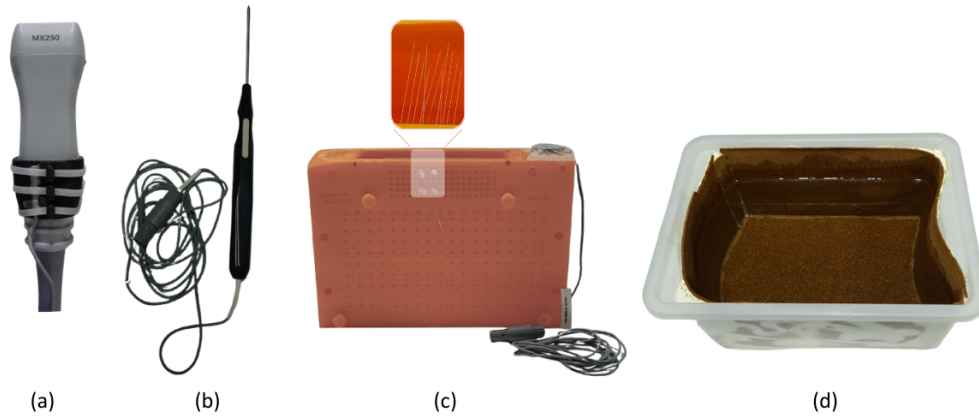


Figure 27: Components used during the calibration procedures: (a) US probe with attached sensor; (b) stylus; (c) calibration phantom; and (d) water-filled container lined with cork to minimize acoustic reflections.

the stylus tip is fixed at a specific point, usually on the phantom itself, ensuring stable positioning within the Aurora FG measurement volume. Using PLUS, the pose of the stylus sensor is recorded while the stylus is moved in circular motion patterns without displacing the tip. This motion allows the software to collect multiple sensor poses while the tip remains stationary. Based on these measurements, the transformation matrix between the stylus sensor coordinate system and the stylus tip coordinate system is estimated by computing the rigid offset using a least-squares approach.

Phantom registration

Following the stylus calibration, the stylus is used to register the reference points printed on both sides of the phantom. This procedure consists of registering twelve points, six on each lateral side, allowing the system to compute the transformation matrix that aligns the virtual phantom coordinate system with the coordinate system of the reference sensor attached to the phantom. This is possible because the software already contains the true positions of these points. By comparing the coordinates acquired by the stylus sensor with the known reference positions defined in the XML file, and using the transformation matrices between the stylus sensor and the reference sensor (both detected by the tracker), the system can accurately align the phantom with its reference sensor. The phantom and its lateral sides containing the reference points are shown in Figure 28.

Temporal calibration

As mentioned above, the two systems differ not only in their acquisition rates, with the MTS operating at 40 Hz and the data at 25 fps, but also in the latency of each measurement, *i.e.* the time required for

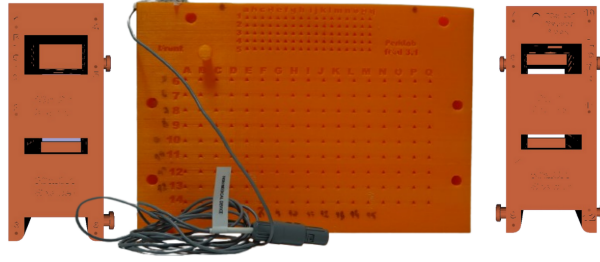


Figure 28: Lateral views of the phantom illustrating the printed reference points employed in the registration step.

a sample to be transmitted to and recorded by the computer. Therefore, it is essential to determine the time offset between the two data streams to correctly align and synchronize them. This offset depends not only on the different acquisition rates and latencies of the systems but can also be affected by computer hardware factors, such as CPU load [79]. Consequently, calibration is performed with minimal background processes running on the computer, both during the calibration procedure and during subsequent data acquisition.

The calibration protocol consists of moving the probe up and down inside a water-filled container. Prior to the motion, the software requires one data stream to be defined as fixed and the other as moving, with the faster channel selected as the reference and the slower channel temporally aligned to it. The movement must be as periodic as possible, and the bottom of the container must appear in the US image, as this serves as a reference point for the algorithm. During acquisition, the bottom of the container is detected in the US images using image processing, and its Y-coordinate is extracted as a signal over time. Similarly, the probe position provided by the MTS is represented as a temporal signal. These two signals, which exhibit sinusoidal waveforms due to the repetitive motion, are then compared using signal processing techniques to determine the time offset corresponding to the highest correlation between signals [79]. To ensure a reliable estimation of the time offset, the temporal calibration is repeated ten times under identical conditions, and the final offset value used in all subsequent acquisitions is defined as the mean of the computed offsets.

Spatial calibration

Finally, the last calibration step determines the transformation matrix that maps pixel coordinates from the image coordinate system to the sensor coordinate system attached to the probe. For this purpose, a scan is performed along the length of the wires placed in the phantom, allowing a comparison between the wire positions defined in the XML configuration file and those obtained from the sensor data. It is essential that the wires remain fully stretched during this procedure, as the calibration algorithm relies on

the accurate detection of their positions in the US images.

To improve the reliability of wire segmentation and intersection detection, several key parameters were empirically adjusted in the calibration configuration. The pixel to millimeter conversion factor, which aims to determine the pixel spacing, should ideally match the resolution of the acquired image. However, due to software limitations, the lowest empirically determined value, 0.0589 mm/pixel, was defined to ensure spatial consistency between the image and tracker domains. Morphological filtering parameters, such as the opening radius (0.2 mm) and bar size (8.6 mm), were tuned to suppress image noise while preserving the wire structures, which have a diameter of approximately 0.3 mm. The binarization threshold was set to 29% of the maximum image intensity, and the maximum distance allowed to fitting collinear points was limited to 0.5 mm. These settings were experimentally tuned to balance detection sensitivity and geometric precision, ensuring robust wire identification and accurate estimation of the spatial calibration matrix. As a result of this calibration, two matrices are obtained: one that transforms the image coordinate system into the sensor coordinate system, and another that defines the pixel to millimeter scaling. For freehand acquisitions, only the first matrix is used, since it already incorporates the pixel to millimeter scaling determined during the calibration process.

To account for possible variability in the calibration procedure, the spatial calibration is repeated ten times, producing ten independent calibration matrices. Their performance is evaluated using the validation setup (described in Section 3.2.1), allowing the selection of the matrix that best represents the *ImageToProbe* transformation.

All coordinate systems referenced throughout this section are summarized in Figure 29.

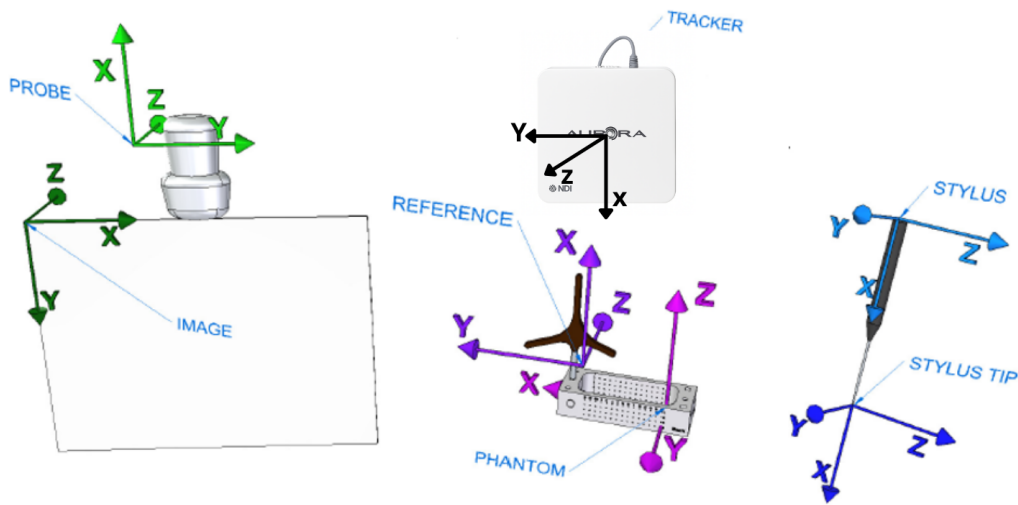


Figure 29: Coordinate systems of the components involved in the calibration process. Adapted from [78].

3.1.3 Data acquisition and preprocessing

After completing all calibration steps, freehand US acquisitions can be performed. In this setup, image and pose data are transmitted to the computer in real time for simultaneous recording. A dedicated XML configuration file is created, specifying the MTS setup, including both sensors, *i.e.* the one attached to the probe and the one tracking involuntary hand movements. The file also incorporates the final spatial calibration matrix, the temporal offset, and the PlusOpenIGTLinkServer settings defining the channels and transformation matrices to be transmitted. Based on this configuration, two types of messages are defined for transmission: image and transform. The transmitted transformations include *ProbeToHandSensor*, *ImageToHandSensor*, *HandSensorToTracker*, *ImageToTracker*, and *ProbeToTracker*. During an acquisition, the PLUS Server Launcher is executed with this configuration, streaming the data in real time to the 3D Slicer software via OpenIGTLinkIF. The overall acquisition and preprocessing workflow is illustrated in Figure 30.

Once the data is received in the 3D Slicer, it is possible to visualize, in real time, the acquired images moving within the 3D space as the probe is moved during acquisition. Using the *Sequence module*, the desired transforms and data streams to be recorded are defined, enabling subsequent preprocessing and volumetric reconstruction.

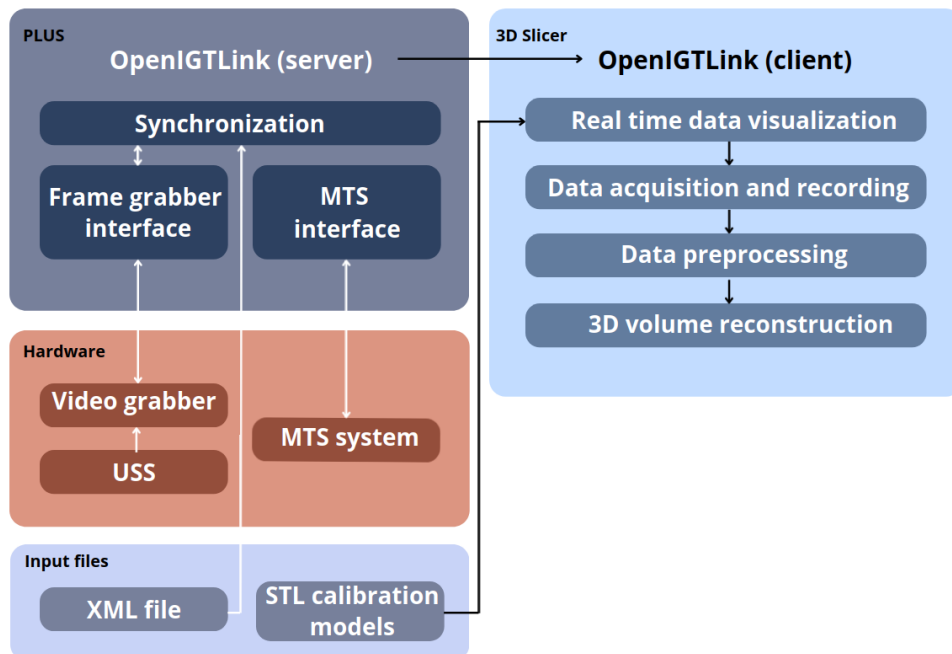


Figure 30: Data acquisition and preprocessing workflow linking the hardware setup, the PLUS server, and the 3D Slicer client through the OpenIGTLink protocol. Adapted from [80].

Preprocessing pipeline

Once the scans are acquired, a preprocessing pipeline is applied to retain only the data relevant for analysis. When a frame grabber captures the US machine's screen output, the images contain additional elements, such as acquisition parameters, US scale markers, and other on-screen information, that are not relevant to the reconstruction, and therefore require cropping. In addition, acquisitions frequently contain unnecessary frames, so a frame range is defined for each acquisition to exclude frames that do not capture the object of interest.

Furthermore, since the spatial calibration was performed for a depth of 30 mm, all matrices derived from this procedure are exclusively valid at this depth. As discussed earlier, the imaging depth and width can be adjusted during acquisition to ensure optimal visualization of the pathology of interest. Consequently, the calibration matrices must be adapted to account for these changes.

As part of the preprocessing pipeline, each acquisition is opened individually in the Napari viewer. During this process, the first and last frames to be retained are specified to ensure that only the relevant portion of the sequence is preserved, and four points are manually selected in the first frame to define the ROI. Two points mark the beginning and end of the image's width, while the remaining two correspond to the upper right and lower right corners of the image, where the scale is displayed. Based on this information, the distance in millimeters indicated by the scale and the corresponding pixel distance measured by the software are used to compute the new pixel resolution. The selected point coordinates are subsequently provided to a Python script, which crops all frames in the sequence, removing the unnecessary information present in the US system screen. Likewise, all pose transformation matrices acquired over time are updated to incorporate the effects of the image cropping operation and the new pixel resolution resulting from changes in the imaging depth, as described in the following paragraphs.

Since the US system automatically performs proportional resizing when adjusting either the depth or width settings, the image maintains its aspect ratio and isotropic scaling. This ensures that the pixel spacing remains identical in both X and Y directions, preventing distortion by stretching or compressing along a single axis. Thus, only the depth parameter had to be considered in the matrix adaptation process.

When the imaging depth is adjusted from the calibration depth of 30 mm, the origin of the cropped US image effectively shifts due to the proportional scaling performed by the system. However, the center of the image, corresponding to the probe's physical center, remains unchanged. To preserve spatial alignment during reconstruction, the original calibration matrix must therefore be post multiplied by an additional transformation, denoted as ${}^{\text{img_calib}}T_{\text{img_ROI}}$, which accounts for both the new imaging depth and the image crop.

This transformation is constructed using two translation matrices that ensure scaling operations are performed about the image center rather than the image origin, thereby preventing undesired shifts in the reconstructed geometry. The first matrix, ${}^{\text{img_calib}}T_{\text{center}}$, shifts the coordinate system from the origin of the calibration image to its center. This formulation is required because the calibration matrix is post multiplied and maps coordinates expressed at the origin of the image space to the probe coordinate system. The second matrix, ${}^{\text{center}}T_{\text{img_ROI}}$, performs the inverse operation for the current acquisition, shifting the coordinate system from the image center to the origin of the cropped ROI image.

The complete transformation applied to adapt the calibration matrix to the new acquisition depth is expressed as:

$${}^{\text{img_calib}}T_{\text{img_ROI}} = {}^{\text{img_calib}}T_{\text{center}} \cdot {}^{\text{px_calib}}T_{\text{mm}} \cdot {}^{\text{mm}}T_{\text{px_ROI}} \cdot {}^{\text{center}}T_{\text{img_ROI}}, \quad (3.1)$$

where ${}^{\text{px_calib}}T_{\text{mm}}$ corresponds to the scaling matrix derived from the calibration at 30 mm (pixels to millimeters), and ${}^{\text{mm}}T_{\text{px_ROI}}$ represents the inverse scaling defined by the new pixel spacing computed for the current acquisition depth.

3.1.4 Volume reconstruction

Following preprocessing, during which the images are cropped, the frame range is selected and the transformation matrices are adjusted (with those corresponding to the discarded frames removed), both image and pose data are loaded into 3D Slicer to perform volume reconstruction.

Within the volume reconstruction module, a ROI is defined to encompass the acquired scan in the 3D space, together with the output spacing and interpolation mode. The output voxel spacing was set to $(0.1 \times 0.1 \times 0.05) \text{ mm}^3$ in the X , Y , and Z directions, respectively. These values were chosen to approximate the average spatial resolution across all acquisitions, as the depth of the US images varies throughout the scans. The interpolation mode is set to linear interpolation, because it produces smoother and visually cleaner reconstructions compared to nearest-neighbor interpolation. Once these parameters are configured, the reconstruction is performed. Finally, the defined ROI and the reconstructed volume are saved for subsequent visualization or analysis.

3.2 Experiments

To evaluate the robustness of the proposed MTS-based reconstruction framework, two complementary experiments were designed. The first involves a tracker-US validation to examine the influence of potential

interferences, such as electromagnetic disturbances or calibration inaccuracies on the spatial accuracy of the Aurora MTS. The second aims to assess the complete pipeline in a more realistic small animal imaging scenario, thereby evaluating its performance under conditions that approximate in vivo acquisitions.

3.2.1 Tracker-ultrasound validation

This validation setup is primarily designed to assess the precision of sensor detection within the field generated by the Aurora system, as well as to evaluate the accuracy of the performed calibration by identifying persistent errors and assessing their impact on the 3D spatial positioning of the acquired images.

To perform the validation, a custom model was developed with six fixed positions to hold the US probe, three on each side, as illustrated in Figure 31. The model was designed to closely match the probe's geometry, ensuring a stable and reproducible fit during measurements. Opposite positions were aligned at equal angles relative to the base to maintain a consistent probe orientation and to ensure that the acquired planes are coincident. The cavities were shaped to match the probe's array, and nine wires were arranged linearly along the model to allow the acquisition of images with these wires visible from each position.

For easier identification during the experiments, one side of the model was labeled with even numbers and the opposite side was labeled with odd numbers. Each position on the same side is separated by a distance of 70 mm. During acquisition, the probe was placed sequentially in each holder and securely fixed with cable ties to minimize any movement. In addition to the sensor attached to the probe, another sensor was placed on the acrylic support to serve as a reference frame. This allows all obtained poses to be expressed relative to a stable coordinate system, thus compensating for small movements of the container or the model itself during handling. It is important to note that the acrylic structure holding the model was firmly fixed to the container using four clamps, ensuring full stability during acquisition. Tests were performed with the model placed inside a water-filled container, allowing proper visualization of the

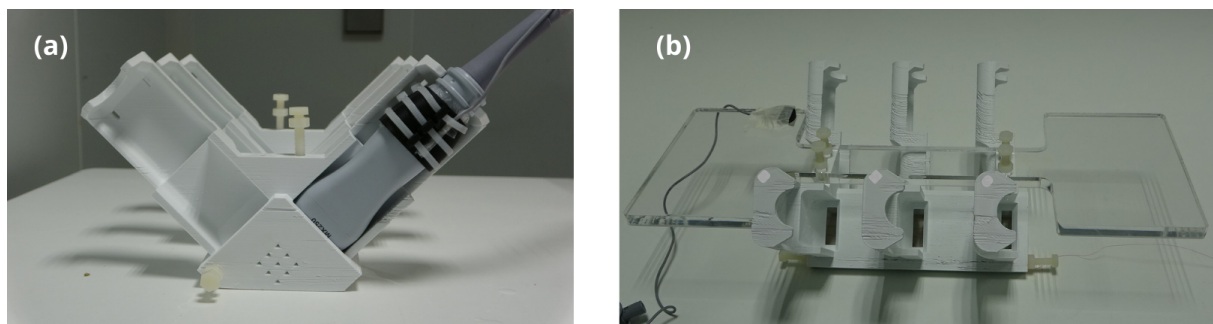


Figure 31: Validation setup showing: (a) the validation support and (b) the acrylic assembly.

wires along the entire structure. The model was designed in SolidWorks and subsequently 3D printed. The complete design and technical drawings are presented in Appendix B.

The acquired videos, obtained in a similar manner to the tumor acquisition sequences, were recorded using the PLUS Server to transmit in real time data, including image and pose information, to the 3D Slicer software running on the computer. During the experiments, the model was positioned both parallel and perpendicular to the FG, as shown in Figure 32. These configurations are designed to evaluate how the system's precision and accuracy vary with the distance from the tracker.

To quantify the system performance, four complementary metrics were computed: positional precision (Pre_p), orientation precision (Pre_o), positional accuracy (Acc_p), and orientation accuracy (Acc_o). These metrics allow the assessment of both the temporal stability of the measurements and their agreement with known GT displacements.

For each static probe position i , 10-second videos were acquired. The positional precision (Pre_p) for each static pose is computed as the standard deviation of the measured positions p_{ij} , with $j = 1, 2, \dots, N_s$, relative to their mean position $\bar{p}_i$, denoted in Equation 3.3. Accordingly, Pre_p is defined as:

$$\text{Pre}_p = \sqrt{\frac{1}{N_s} \sum_{j=1}^{N_s} \|p_{ij} - \bar{p}_i\|_2^2}. \quad (3.2)$$

$$\bar{p}_i = \frac{1}{N_s} \sum_{j=1}^{N_s} p_{ij}, \quad \bar{p}_i \in \mathbb{R}^3, \quad (3.3)$$

where $\bar{p}_i$ is obtained as the arithmetic mean of all translation vectors recorded during the 10-second acquisition for pose i .

Likewise, orientation precision (Pre_o) is determined from the distribution of the orientation quaternions q_{ij} , expressed as:

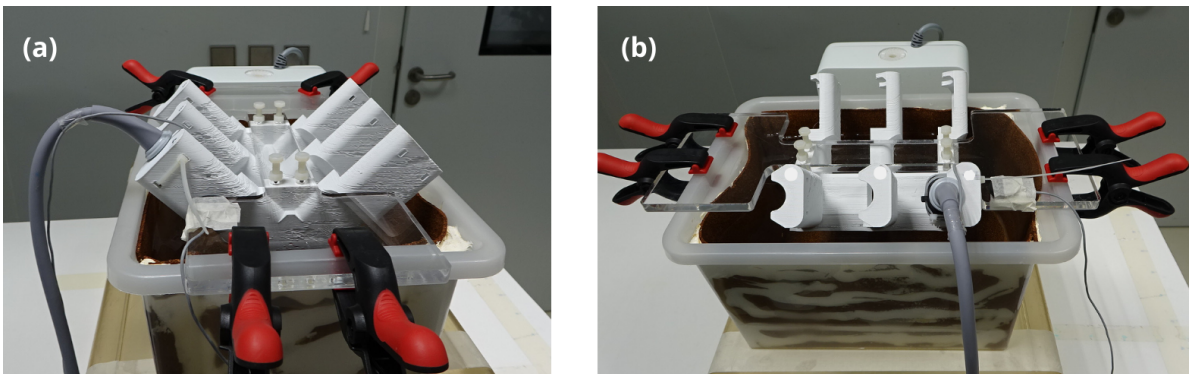


Figure 32: Illustration of the experimental setups used in the validation: (a) perpendicular configuration and (b) parallel configuration.

$$\text{Pre}_0 = \sqrt{\frac{1}{N_s} \sum_{j=1}^{N_s} (2 \arccos(q_{ij} \bar{q}_i^*))^2}, \quad (3.4)$$

where q_{ij} represents the quaternion describing the orientation of the j -th frame and $\bar{q}_i$ denotes the mean quaternion of all samples acquired during the 10-second interval corresponding to that static position. The mean rotation $\bar{q}_i$ is computed using the quaternion-based averaging scheme proposed by Markley *et al.* [81], which is described below. The term $(\cdot)^*$ denotes the quaternion conjugate, corresponding to the inverse rotation, and is used to compute the relative angular difference between two orientations.

To determine $\bar{q}_i$, the *Markley* averaging method constructs a symmetric matrix M from all quaternions:

$$M = \sum_{j=1}^{N_s} w_j q_j q_j^T, \quad (3.5)$$

with optional weights w_j (set to 1 in this work). The mean quaternion is obtained as the unit eigenvector associated with the highest eigenvalue of M , which can equivalently be written as the maximization problem given by:

$$\bar{q}_i = \arg \max_{q \in \mathbb{S}^3} q^T M q, \quad (3.6)$$

where $\mathbb{S}^3$ denotes the unit 3D sphere of normalized quaternions. This formulation yields a statistically robust estimate of the average rotation while ensuring the resulting quaternion remains properly normalized.

The Acc_p is computed as the absolute difference between the GT distance (d_{GT}) and the measured Euclidean distance ($\bar{d}_i$) between two probe locations, according to:

$$\text{Acc}_p = |d_{GT} - \bar{d}_i|, \quad \text{with} \quad \bar{d}_i = \|\bar{p}_i - \bar{p}_{i+2}\|_2, \quad (3.7)$$

The analysis is performed for pairs of positions on the same side of the model (positions 1–3, 3–5, 2–4, and 4–6), each separated by a GT distance of 70 mm. This metric quantifies the tracking accuracy when the probe is displaced by a known distance.

Similarly, the *orientation accuracy* (Acc_o) is defined as the absolute difference between the GT angular displacement (α_{GT}) and the measured mean rotation angle ($\bar{\alpha}_i$) between two probe orientations, according to:

$$\text{Acc}_o = |\alpha_{GT} - \bar{\alpha}_i|, \quad \text{with} \quad \bar{\alpha}_i = 2 \arccos(\bar{q}_i \bar{q}_{i+2}^*), \quad (3.8)$$

where $\bar{q}_i$ and $\bar{q}_{i+2}$ represent the mean quaternions corresponding to two adjacent probe positions on the same side of the model, obtained from the 10-second static acquisitions. As in the case of Acc_p , the same pairs of positions were considered, each separated by a known GT angular displacement of 0° , since the custom model was designed to hold the probe at the same orientation across all positions.

To assess how tracking errors propagate to the image coordinate system, the wires visible in each US image acquired at the six model positions are segmented to determine the corresponding centroids. The images are first cropped in the same way as described in the previous subsection and binarized using a fixed intensity threshold. After binarization, the small components are then removed through morphological filtering, and connected components are identified. For each region, the centroid is refined within a local mask by computing the intensity weighted center of mass after Gaussian smoothing, yielding subpixel centroid coordinates.

To obtain a consistent set of targets, a 3×3 grid of wire centroids is selected by clustering the coordinates along the vertical and horizontal axes, based on the expected grid like arrangement of the wires. This procedure enforces the selection of nine valid points while ensuring the removal of false positives.

Before spatial mapping, the centroid coordinates are converted to homogeneous form. Let $p_k = [x_k, y_k, 0, 1]^T$ denote the homogeneous coordinate vector of the k -th centroid in the image coordinate system, with $k = 1, \dots, 9$. Each centroid is then transformed into the chosen reference frame, either the tracker or the acrylic sensor using the averaged pose for that acquisition, namely the image to tracker (${}^T T_{\text{img}}$) or image to acrylic (${}^A T_{\text{img}}$) matrix, according to $p_k^{\text{ref}} = T [x_k \ y_k \ 0 \ 1]^T$, where p_k^{ref} represents the position of the k -th wire centroid expressed in the chosen reference coordinate system.

In an ideal system free from both random and systematic errors, the points representing the wires in images acquired from opposite positions would perfectly overlap when registered in 3D space. However, as previously discussed, small inaccuracies are always present, and points that should be collinear in practice deviate from an ideal straight line. To quantify this deviation, the corresponding sets of wire points (one point for each acquisition position) are identified and a straight line is fitted to each set using a principal component analysis algorithm. This algorithm determines the line direction that maximizes the variance of the data and best represents the underlying spatial alignment of the points. The orthogonal distances of each point to the fitted line are then calculated, providing an estimate of the spatial deviation.

A dispersion measure was obtained by computing the mean orthogonal distance to the fitted line, denoted as MAD, defined as:

$$\text{MAD} = \frac{1}{N} \sum_{k=1}^N d_k, \quad (3.9)$$

where d_k represents the orthogonal distance between point k and the best fit line, and N is the total number of points.

3.2.2 Volume reconstruction validation

After performing volume reconstruction based on the pose data provided by the MTS, validating the results is essential. To this end, a series of scans were acquired and recorded, followed by preprocessing and 3D reconstruction using 3D Slicer. To validate the reconstructed volumes, the mechanical scanning system of the Vevo 3100 US machine was used to acquire the same target data. This enabled a direct comparison between volumes reconstructed from freehand acquisitions and those obtained using the mechanical system, providing a reference for assessing the accuracy and reliability of the MTS-based freehand reconstruction pipeline.

Dataset

Initially, the aim was to acquire tumor images and perform tumor volume reconstruction in mice. However, due to the limited availability of research studies involving tumors in mice at the ICVS animal facilities during the acquisition period, the protocol was adapted to focus on a specific anatomical organ. In a first instance, two anatomical structures were considered as potential imaging targets: the kidneys and the bladder, as these are among the most clearly visible organs in abdominal US imaging. The bladder only approaches the size of the kidney when fully distended, reaching approximately 8–9 mm in diameter, and its volume is highly variable depending on the degree of filling. This variability would be difficult to control during acquisitions and would introduce a significant confounding factor into the dataset. For these reasons, the kidney was selected. In addition, each mouse possesses two kidneys, which enables the acquisition of a larger and more diverse dataset without increasing the number of animals used. It is also important to note that, although relatively large compared to other murine organs, the kidney remains smaller than typical tumors.

Moreover, although it was initially envisioned that acquisitions would be performed on awake mice, restrained using techniques commonly employed by researchers, the lengthy authorization process required for experiments on live animals required that the experimental scans in this thesis be conducted on animals that had already been euthanized.

For image acquisition, the dorsal region of the animal was first shaved, followed by the application of a depilatory cream to carefully remove all remaining fine hairs, thereby avoiding compromising the US image quality. The area was subsequently cleaned with alcohol to remove any residual cream and to ensure that the region was thoroughly prepared for imaging.

Scanning protocol

Prior to initiating the acquisitions, a scanning protocol was defined to ensure that the resulting images were consistent and suitable for dataset creation. The protocol consisted of freehand scans, allowing the researcher to move the probe as needed while visualizing the target anatomy. Both transverse and longitudinal scans of the left and right kidneys were acquired. During transverse acquisitions, the probe was moved in the superior-to-inferior direction, *i.e.*, from the head toward the hind limbs. In longitudinal acquisitions, the movement was performed from the abdomen toward the spine.

Each acquisition began with positioning the mouse in the researcher's hand, along with the placement of the sensor used to monitor undesired hand movements relative to the Aurora MTS. Image acquisitions were then performed using the tracked US probe. As is common in this type of acquisition, the probe was moved slowly, with total scan durations ranging from 5 to 10 seconds, depending on the imaging plane being acquired. The Vevo system parameters were set as follows: US frequency of 21 MHz, power at 100%, frame rate of 58 fps, gain of 24 dB, dynamic range of 60 dB, display map set to G5, Vevo HD enabled, and B-mode selected. As previously described, the image depth and width were not fixed, allowing the researcher to adjust both to better focus on the anatomical structure of interest.

For acquisitions performed using the mechanical scanning system, mice were positioned in the prone orientation, with all limbs secured to the platform to ensure immobilization during scanning. A probe movement step size of 0.04 mm was defined, while depth and width parameters were set similarly to those used in freehand scanning. The scan range was selected to ensure complete coverage of the kidney.

Evaluation metric

Upon acquisition, the kidneys were segmented from both freehand and mechanical reconstructed volumes. All segmentations were performed by the same operator to ensure consistency. The freehand reconstructions were segmented using the 3D Slicer software, whereas mechanical scans were segmented in the VevoLAB software, as these data were only available in Vevo's native file format.

To assess the pipeline's accuracy, the volumes of the segmented object in both reconstructions were compared using the RVD metric described in Section 2.4.2.

3.3 Results and discussion

3.3.1 Tracker-ultrasound validation

Calibration results

In the stylus calibration, the acquisition of 400 points was specified in the XML configuration file, resulting in an error of 0.341 mm. It is important to understand that this error does not necessarily indicate poor calibration quality or system malfunction. In particular, it was observed that reducing the number of acquired points or constraining the range of stylus movement resulted in lower error values. This behavior can be explained by the fact that fewer points or a more limited movement range reduce pose variability, thereby introducing fewer numerical errors into the estimation. However, using a larger number of points and a wider spatial coverage provides a more representative sampling of the tracker's operating volume and thus produces a calibration that is more reliable for realistic conditions and for registering the reference points of the phantom, even if the reported error appears higher.

For the phantom registration, an error of 0.893 mm was obtained, which is within the expected accuracy range for electromagnetic MTSs. Such a low registration error demonstrates that the mapping between the measured fiducial positions and their corresponding reference points was accurately established, ensuring that the transformation between the phantom and tracker coordinate systems is reliable.

In the temporal calibration, the obtained offset values ranged from 0.096 s to 0.121 s, resulting in a mean offset of 0.112 s. This value represents the time delay between the acquisition of the US images and the corresponding tracker data. It is essential that this offset closely matches the true delay, as a mismatch would cause the software to associate poses with incorrect frames, leading to spatial misalignment, reconstruction artifacts, and reduced geometric precision in the resulting volume.

Finally, for spatial calibration, and considering that the error obtained at this stage depends on the matrices from the previous calibration steps, an error range between 0.819 mm and 1.290 mm was obtained across the ten acquired calibration matrices. It is also important to note, as mentioned earlier for the stylus calibration, that a lower numerical calibration error does not necessarily imply a better spatial calibration. Smaller errors may arise from a reduced range of probe motion or a limited sampling of poses, which can make the optimization problem easier but less representative of realistic operating conditions.

Therefore, a dedicated validation setup (described in Section 3.2.1) was used to quantitatively evaluate the performance of the obtained calibration matrices and identify the one that best represents the true *ImageToProbe* transformation. The detailed results of this assessment are presented later in this section.

Tracker-ultrasound validation setup

The results presented in Table 1 summarize the MTS' precision and accuracy for both parallel and perpendicular setups. Prior to summarization, a *Shapiro–Wilk* normality test was performed to determine the appropriate statistical representation of the data. Since some of the measured values did not follow a normal distribution ($p < 0.05$), all results are reported as median [Q1, Q3] to ensure consistency and robust comparison between metrics. As mentioned in Section 3.2.1, two reference systems could be considered to express the transformation matrices: the tracker coordinate system or the coordinate system of the sensor attached to the acrylic platform. The latter option would account for any potential undesired movements of the platform during acquisition. However, after analyzing the data referenced to both systems, it was observed that the acrylic support remained highly stable throughout all acquisitions, showing negligible motion in every position. For this reason, all results were expressed relative to the tracker reference system, thus avoiding the introduction of an additional transformation step that could propagate compounding errors.

The results obtained for precision and accuracy in the sensor attached to the probe demonstrated highly acceptable values, considering the known limitations of the Aurora system, and are consistent with other evaluation studies reported in the literature [82]. To further investigate how precision and accuracy vary with the distance from the tracker, the values obtained for Pre_p , Pre_o , Acc_p , and Acc_o at each of the six positions were plotted, as shown in Figure 33 and 34, respectively.

When observing the 3D spatial distribution of the sensor positions relative to the FG origin, it can be noted that the adjacent pairs of positions, (1, 2), (3, 4), and (5, 6), are not perfectly symmetric along the Z axis, with the even numbered positions appearing slightly farther from the tracker. This effect arises from the sensor being mounted on the lateral side of the US probe rather than in its central plane. Additionally, the probe was rotated by 180° when acquiring positions on opposite sides, which contributes to the observed displacement pattern. In theory, the spatial calibration matrix fully compensates for the probe geometry, so this offset should not affect the reconstructed image plane or any measurements derived from it.

For both positional and orientation errors, it was observed that as the distance from the tracker in-

Table 1: Precision and accuracy metrics for the system under the parallel and perpendicular setups (median [Q1–Q3]).

	Position precision [mm]	Orientation precision [°]	Position accuracy [mm]	Orientation accuracy [°]
Parallel	0.012 [0.005 – 0.019]	0.015 [0.005 – 0.032]	0.417 [0.295 – 0.529]	0.937 [0.363 – 1.637]
Perpendicular	0.012 [0.007 – 0.020]	0.010 [0.005 – 0.011]	1.011 [0.603 – 1.326]	0.222 [0.143 – 0.413]

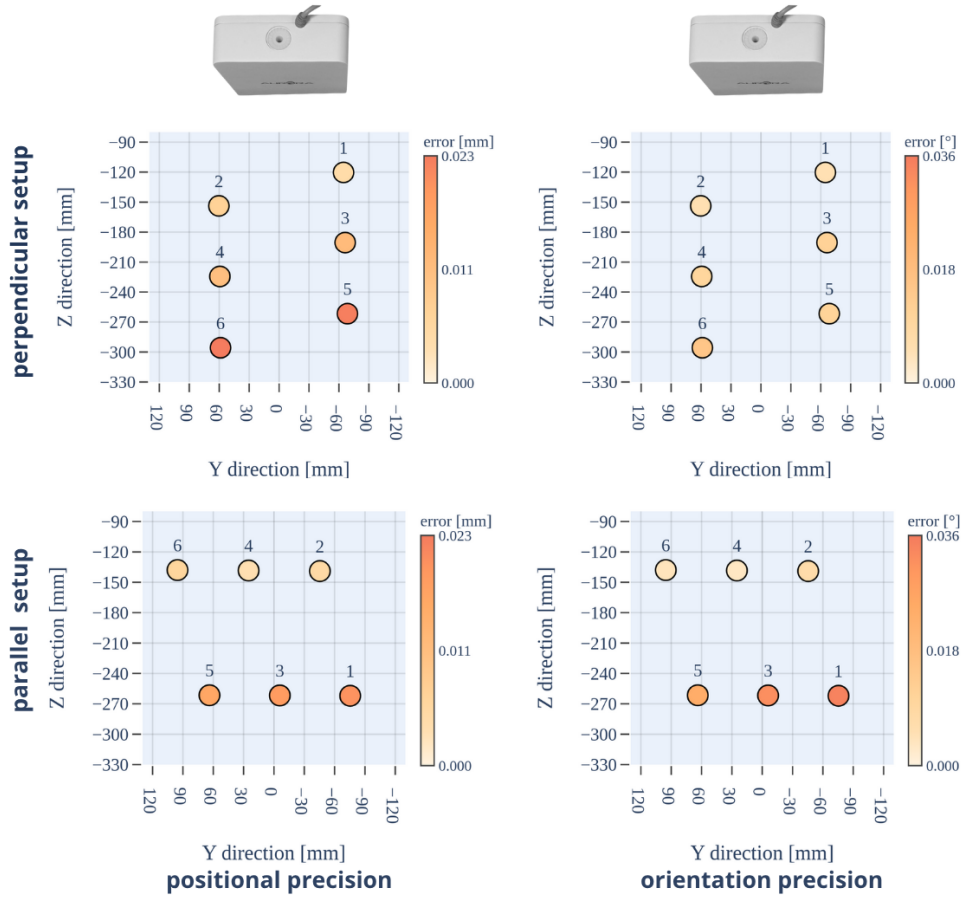


Figure 33: Spatial visualization of positional (left) and orientation (right) precision for both the perpendicular (top) and parallel (bottom) setups.

creases, the corresponding accuracy and precision values also tend to worsen. This trend is consistent with findings reported in the literature [83, 84] and aligns with the operating principles of the Aurora MTS, since the electromagnetic field becomes more susceptible to distortions as the sensor moves farther from the FG. Regarding precision, both positional and orientation jitter remained highly consistent across all probe positions, with only a slight increase observed at greater distances, confirming the high temporal stability of the Aurora system during static acquisitions.

In the accuracy plots, the central points appear in a neutral (gray) color, since the positional and orientation errors were computed relative to these reference positions. This representation clearly illustrates how deviations increased as the probe moved away from the tracker, reflecting the spatial dependency of the electromagnetic field and the gradual degradation of accuracy toward the limits of the tracking volume.

Overall, both parallel and perpendicular configurations yielded comparable results, confirming that the tracking accuracy and stability of the proposed setup were well within acceptable limits for freehand US reconstruction within the defined workspace.

In the MAD metric, the objective was to evaluate how well the points acquired in the six positions align

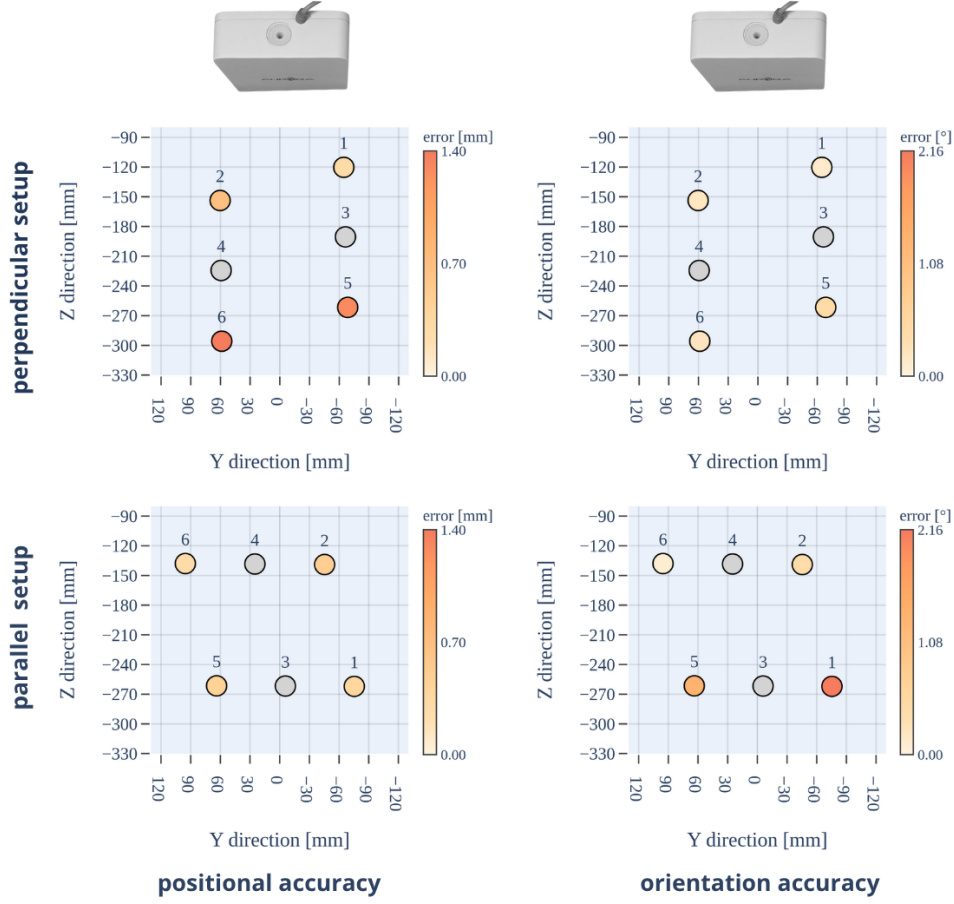


Figure 34: Spatial visualization of positional (left) and orientation (right) accuracy for both the perpendicular (top) and parallel (bottom) setups.

along the corresponding wire, now expressed in the image coordinate system. At this stage, the quality of the *ImageToProbe* (spatial) calibration becomes particularly relevant, since any inaccuracy in this matrix directly affects the computed point coordinates and, consequently, the quality of the line fitting. The results of this analysis are summarized in Table 2.

Among the ten calibration matrices, calibration 8 achieved the lowest MAD, indicating the most consistent mapping between the image and tracker systems. For comparison, the same metric was also computed for an additional calibration obtained by averaging the ten calibration matrices. In this averaged calibration, the rotation component was determined using quaternion averaging, while the translation component was computed as the arithmetic mean of the ten translations, as described in the previous section. Additionally, per-axis scale factors were extracted from each transformation matrix, averaged across all calibrations, and subsequently reintroduced into the final averaged transformation. This averaged matrix exhibited intermediate performance, producing better results than some individual calibrations but worse than others. This behavior can be attributed to the fact that the averaged matrix was not optimized for a specific calibration dataset, instead it represents a general mean that can be influenced by outliers.

Table 2: MAD computed across all calibration sets for the combined parallel and perpendicular setups.

Calibration	PLUS calibration error [mm]	MAD – Median [Q1–Q3]
1	0.908	1.000 [0.887 – 1.143]
2	1.112	0.921 [0.801 – 1.061]
3	0.819	1.185 [1.074 – 1.334]
4	0.900	0.83 [0.725 – 0.956]
5	0.856	0.616 [0.590 – 0.689]
6	1.289	0.522 [0.508 – 0.535]
7	0.985	0.708 [0.614 – 0.816]
8	0.862	0.518 [0.506 – 0.539]
9	1.089	0.646 [0.596 – 0.755]
10	1.119	1.347 [1.203 – 1.539]
Mean calibration	—	0.9189 [0.8076 – 1.0845]

Values represent the aggregated performance (median [Q1–Q3]) obtained for each spatial calibration. The first column lists the calibration error (in millimeters) reported by the PLUS Toolkit during each individual calibration acquisition. The lowest MAD value is highlighted in bold, corresponding to the selected calibration.

Importantly, the calibration error reported by the PLUS Toolkit did not correspond to the lowest MAD value, reinforcing the relevance of this validation setup for selecting the most representative calibration matrix. For this reason, the MAD metric was adopted as the selection criterion, and the calibration corresponding to the lowest MAD (calibration 8) was chosen to be used in any subsequent freehand acquisitions. Overall, the low MAD values obtained validate the reliability of the calibration process and confirm that the selected calibration matrix provides consistent transformations for accurate reconstruction.

In Figure 35, an example of the acquired 3D point distribution and the corresponding best fit lines is shown. The points appear consistently aligned along each wire, confirming the overall accuracy of the calibration and supporting the results obtained from the MAD metric. However, a small gap can be observed between adjacent wire segments from opposite positions in the model that should ideally be collinear and overlapping. This misalignment is likely due to the 180° rotation of the probe, which was not fully compensated by the spatial calibration matrix.

Additionally, the planes corresponding to opposite probe positions were expected to be coincident, implying an angular deviation of 0° between them. To verify this, the angular deviation between adjacent positions was calculated using the dot product of the mean quaternions associated with each position, following the same procedure described above for computing the Acc_o . This analysis was performed using the best calibration, yielding the values presented in Table 3.

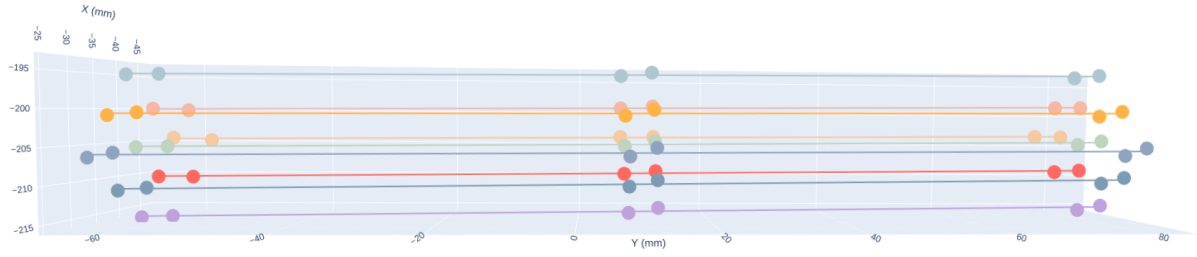


Figure 35: Example of the 3D distribution of the segmented wire points and corresponding best fit lines in the parallel setup, expressed in the image coordinate system.

Table 3: Angular deviation between adjacent probe positions, computed for the best calibration.

Pair	Angular deviation [°]
pos1–pos2	0.990
pos3–pos4	0.244
pos5–pos6	0.373

All measured angular deviations were below 1° , confirming that the probe maintained an almost identical orientation between opposite positions. The results were highly consistent across all calibrations (data not shown), with negligible variation between them.

3.3.2 Volume reconstruction validation

Tracker-based vs mechanical reconstruction

To validate the volume reconstruction based on the poses provided by the MTS, a total of 16 mice were scanned. The animals were aged 2 to 11 months, comprising 9 females and 7 males. For each mouse, both kidneys were recorded, with six acquisitions performed in the longitudinal direction and six in the transverse direction, resulting in a total of 24 scans per mouse. On average, each acquisition lasted between 5 and 10 seconds, corresponding to approximately 100 to 200 frames per acquisition.

Due to time constraints, three mice were randomly selected for the validation experiment. As shown in Table 4, the segmented kidney volumes for these mice are reported for both the freehand and mechanical acquisitions. For each mouse, two independent freehand-derived reconstructions were segmented, yielding similar volume estimates. The relative differences between repeated freehand volumes were 3.07%, 1.81%, and 7.00% for mouse 1, mouse 2, and mouse 3, respectively. These results indicate a low intra-mouse variability, suggesting that the segmentation process was consistently and that the proposed reconstruction pipeline produces repeatable volume estimates across independent freehand acquisitions.

Table 4: Comparison of MTS-based and mechanical-scan volumes

Mouse	MTS-based volume (mm ³)	Mechanical scan volume (mm ³)	RVD (%)
Mouse 1	170.50	161.42	3.86
	165.33		
Mouse 2	441.02	416.32	6.46
	449.09		
Mouse 3	124.31	145.11	12.65
	133.32		

When comparing the freehand volumes with the corresponding mechanical (GT) volumes, the values obtained from the electromagnetic MTS-based reconstructions remain within a reasonable range for all three mice. It is also important to note that the segmentations were performed in two different software platforms (3D Slicer for freehand reconstructions and VevoLAB for mechanical scans), which can introduce additional small discrepancies due to differences in rendering, interpolation, and segmentation tools.

Overall, the freehand-based volumes remain close to the mechanical reference values, indicating that the MTS-based reconstruction is capable of producing kidney volumes of comparable magnitude. To further quantify this agreement, the relative volume difference (RVD) between the two modalities was computed (upon averaging the two estimates of the freehand-based kidney volume). The RVD values ranged from approximately 3.9% to 12.6% across the three mice. Considering the manual segmentation process and the inherent variability of freehand acquisitions setup, deviations of this magnitude can be regarded as acceptable. In Figure 36, a qualitative comparison for the intermediate case (mouse 2) is shown. Both reconstructions appear anatomically consistent with the expected kidney shape and visually similar to each other.

Tumor reconstruction

As noted previously, mice with tumors were not available during the main acquisition period. However, two specimens were later collected, allowing for a preliminary qualitative assessment of tumor reconstruction. Unlike the kidney, the tumor presents clearer and more well defined boundaries, making it easier to visualize and identify in the US images. Figure 37 shows an example of a tumor frame acquired during the scan, together with the corresponding tumor segmentation extracted from the volume reconstructed using the entire sequence.

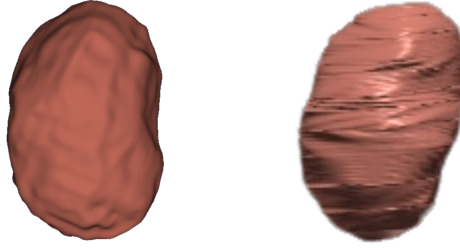


Figure 36: Qualitative comparison of kidney segmentations from MTS-based reconstruction (left) and mechanical scanning (right), for mouse 2, selected as the intermediate case.

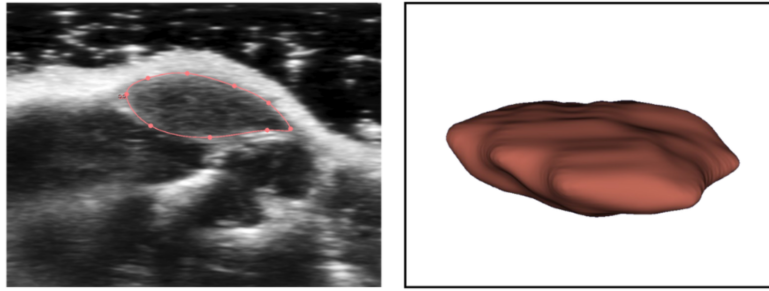


Figure 37: Example of tumor acquisition and reconstruction: (a) acquired US frame showing the tumor with the annotated tumor boundary; and (b) 3D tumor segmentation extracted from the reconstructed volume.

3.4 Conclusion

The validation experiments with the Aurora tracker confirmed that the chosen calibration provided accurate and stable pose measurements, ensuring reliable input for the reconstruction pipeline. The proposed validation setup demonstrated that calibration quality cannot be assessed solely based on the error reported by PLUS, and that an independent spatial validation metric is required.

From the perspective of volumetric reconstruction, the quantitative evaluation demonstrated that the reconstructed kidney volumes closely agree with mechanical ground-truth scans, supporting the reliability and repeatability of the proposed MTS-based approach for organ volume estimation.

The preliminary tumor reconstructions further confirmed that the system performs effectively for the intended use case, supporting its suitability for both anatomical and pathological volume assessment. Nevertheless, it is essential to ensure that the room setup remains as consistent as possible, as electromagnetic tracker-based systems are highly sensitive to external sources of interference. As a consequence, environmental changes may distort the electromagnetic field and affect reconstruction accuracy.

Chapter 4

Trackerless pose estimation

This chapter presents the methodology for performing volume reconstruction using a trackerless approach. As the acquisition of the mouse dataset described in Chapter 3 was still ongoing, the proposed methods were initially developed and evaluated using the publicly available Trackerless 3D Freehand Ultrasound Reconstruction (TUS-REC) challenge dataset [85], which provides large US sequences suitable for early experimentation and model benchmarking. Nevertheless, due to the time constraints associated with the completion of this dissertation and the insufficient number of mice acquisitions available to support model training within this period, the trackerless approaches were not evaluated on the mice dataset. Their assessment in this context therefore remains an open direction for future work.

4.1 Methods

The trackerless approach proposed in this chapter replaces the external electromagnetic tracker with a DL model that estimates the 6-DoF pose of each frame directly from the acquired US sequence.

The first section introduces the overall formulation of the trackerless pose estimation problem and outlines the adopted DL framework.

The second section describes the model architecture selected for this task, EfficientNet-B1, detailing the motivation for its use and the adaptations needed to process US sequences.

The third section presents the different input and output configurations evaluated in this work, including paired-frame and window-based inputs, reference frame and adjacent frame outputs, and the multi-transformation strategy.

The fourth section describes the ensemble strategies explored to improve prediction stability, the training loss functions employed, and the implementation details used to train and evaluate the models.

4.1.1 Overview

In a trackerless freehand US framework, volumetric reconstruction relies entirely on the accurate estimation of probe motion directly from image data, as no external tracking information is available. This problem can be formulated as the estimation of one or more 6-DoF transformations for each frame in an US sequence, describing how individual frames are positioned relative to one another in space.

To improve robustness and stability, the proposed pipeline explores multiple input–output configurations and ensemble strategies, allowing the model to account for variability in motion patterns and prediction. Together, these components form a reconstruction framework in which pose estimation, temporal modeling, and prediction aggregation are combined to enable trackerless volumetric reconstruction from sequential US images. This section provides a conceptual overview of this strategy and establishes the context for the methodological components detailed in the following subsections.

4.1.2 Model architecture

CNNs have been widely used in image classification, segmentation, and many other computer vision tasks due to their ability to learn rich feature representations through the use of convolutional kernels. However, as CNNs become deeper and wider, they often become computationally expensive and difficult to deploy efficiently [86]. To address these limitations, several efficient network families have been proposed in recent years. Among them, the EfficientNet models constitute a particularly successful approach, achieving a good balance between accuracy and computational cost through a principled scaling strategy [87].

The EfficientNet family includes several model variants, ranging from B0 to B7, each representing a different trade-off between accuracy and computational efficiency. These variants share the same architectural principles, but differ in their depth, width, and input resolution, which are jointly scaled using compound scaling. As a result, the accuracy of the EfficientNet models (B0–B7) increases progressively as the number of parameters and floating-point operations increases [87].

A key component of EfficientNet is the use of mobile inverted bottleneck convolution (MBConv) blocks originally introduced in MobileNetV2. These MBConv blocks combine inverted bottlenecks and depthwise separable convolutions, enabling the network to achieve high representational power while significantly reducing the number of parameters and floating-point operations [86].

Figure 38 illustrates the block level structure of EfficientNet-B1, the backbone selected for the trackerless pose estimation approach in this work. The architecture is composed of an initial 3×3 convolution

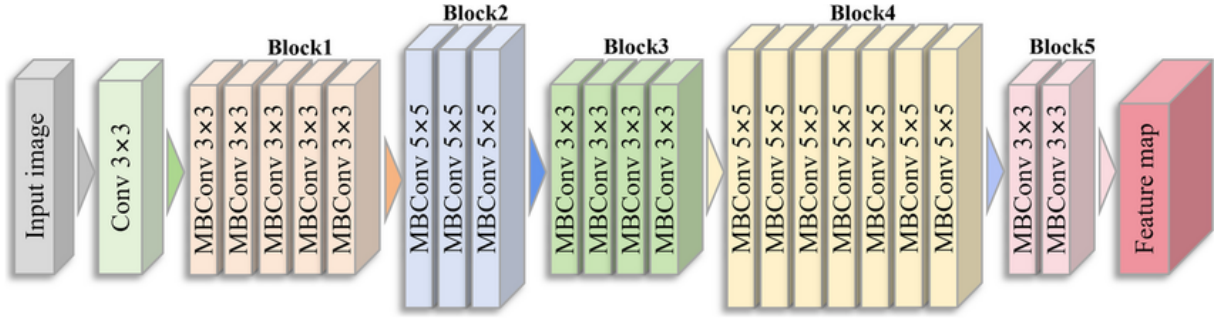


Figure 38: Overview of the EfficientNet-B1 architecture. From [88].

followed by a sequence of MBConv blocks with varying kernel sizes and depths. These blocks combine depthwise convolutions with lightweight pointwise convolutions, enabling the extraction of rich feature representations while keeping computational cost and memory requirements low. This architecture is also well suited for preserving fine-grained spatial details [87], an essential property for improving the accuracy of pose estimation.

4.1.3 Input variants

Conventionally, CNNs process RGB images represented by three channels corresponding to color information. In this work, instead of color channels, the input channels encode temporal information from consecutive US frames (each with one channel only). Each frame is treated as an independent channel and stacked along the channel dimension, allowing the network to learn spatial-temporal features directly through standard 2D convolutions. This design enables the network to implicitly capture motion information without relying on explicit temporal architectures such as RNNs or 3D convolutions.

Within this channel-stacking formulation, different strategies for selecting and grouping frames lead to distinct input configurations. Specifically, two input variants are considered.

The first strategy, termed paired-frame input, uses two consecutive frames selected from the scan and stacked along the channel dimension. These frames are provided as input to the network, which is trained to predict the 6-DoF relative pose transformation between them.

The second strategy, referred to as the window-based input, considers a larger temporal context. A window of N consecutive frames is stacked along the channel dimension, forming an N channel input. In this configuration, two window sizes are evaluated, $N = 10$ and $N = 100$, allowing the model to access richer temporal information that can improve the estimation of frame-to-frame motion. Both approaches are illustrated in Figure 39.

Another factor investigated is the effect of image resolution, which can influence pose prediction

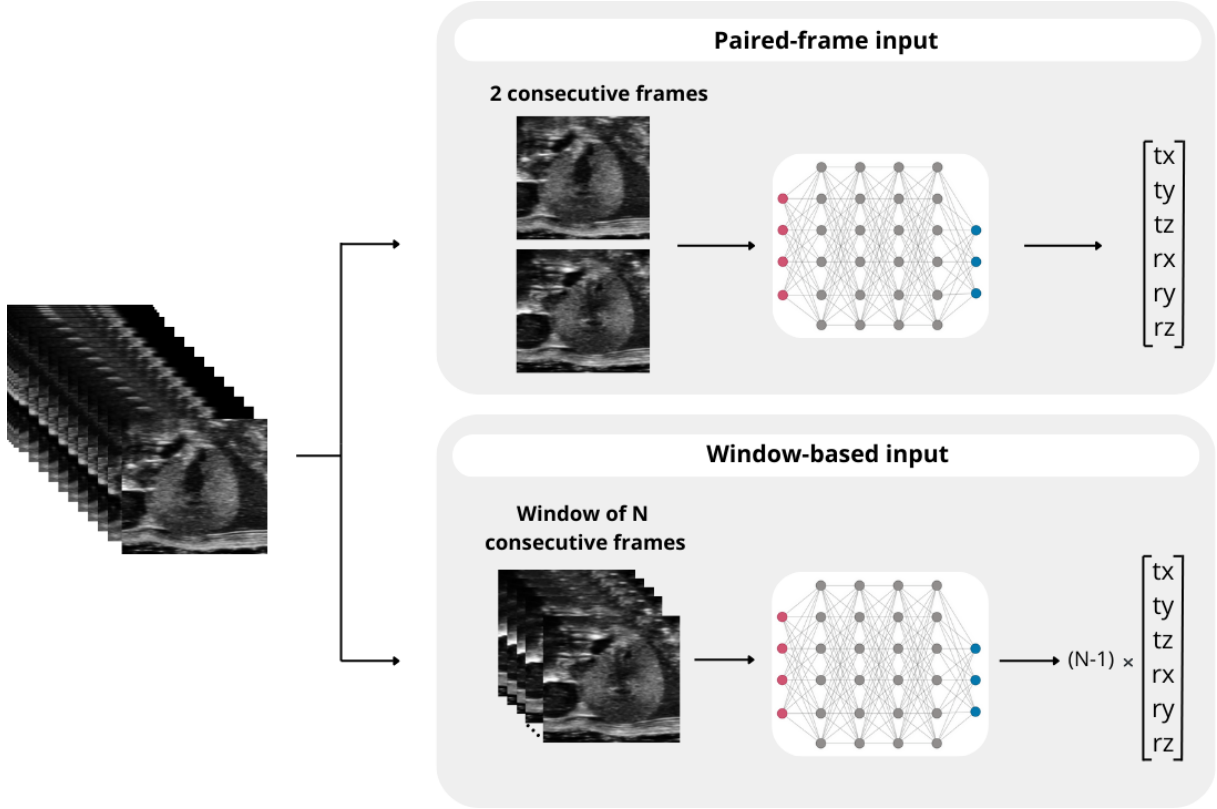


Figure 39: Illustration of the paired-frame and window-based input strategies for pose estimation.

accuracy as previously mentioned. For this purpose, models are trained using both the original image resolution and a downsampled version obtained with a factor of 4. Downsampling is performed with bilinear interpolation (order=1) using the scikit-image library, with anti-aliasing enabled to reduce potential artifacts. This bilinear interpolation computes the output value as a weighted average of neighboring pixels, producing smoother results than nearest-neighbor interpolation.

4.1.4 Output variants

Depending on the input configuration provided to the network, different output variants are also evaluated. A relevant aspect is that the network predicts 6-DoF motion parameters, which are subsequently converted into homogeneous transformation matrices using PyTorch functions. These predicted matrices are then compared with the GT transformations, which are provided in the dataset as 4×4 homogeneous matrices expressed in the world coordinate system (the tracker's coordinate system).

Adjacent frame transformation

Adjacent frame transformations correspond to the estimation of the pose between consecutive frames and can be performed using either the paired-frame input or the window-based input.

For the paired-frame input configuration, the network is trained to predict the relative motion between two consecutive frames. Accordingly, in this inference mode, the prediction sequence begins with the first pair of frames ($\text{frame}_0, \text{frame}_1$) to estimate the initial local transformation. A sliding window with stride 1 is then applied, such that each frame is compared to its immediate predecessor. To evaluate the global performance, the locally predicted transformations are recursively accumulated along the sequence, as shown in Equation 4.1:

$${}^0T_i = {}^0T_1 \cdot (\dots) \cdot {}^{i-1}T_i, \quad (4.1)$$

where ${}^{i-1}T_i$ denotes the transformation from frame i to frame $i - 1$.

Since the GT poses provided in the dataset are expressed in world coordinates, and the loss is computed with respect to the previous frame, the labels need to be converted into frame-to-frame transformations. This conversion follows Equation 4.2, where ${}^wT_i^{\text{GT}}$ denotes the pose of frame i in the world coordinate system w :

$${}^{i-1}T_i^{\text{GT}} = ({}^wT_{i-1}^{\text{GT}})^{-1} {}^wT_i^{\text{GT}}, \quad i = 1, \dots, N - 1, \quad (4.2)$$

where ${}^{i-1}T_i^{\text{GT}}$ denotes the transformation that maps points from the coordinate system of frame i to that of frame $i - 1$.

For the window-based input, although a set of N frames is fed into the model, the network is trained to predict frame-to-frame (local) pose transformations, i.e., the pose of each frame relative to its immediate predecessor. In this configuration, inference produces $N - 1$ local pose predictions for a window of N frames, corresponding to transformations between consecutive frames.

The inference procedure follows the same general structure as in the paired-frame variant, with an additional accumulation step. Since the network outputs local transformations within each window, these predictions are recursively accumulated within window k to obtain the pose of each frame relative to the coordinate system of the first frame in that window. This accumulation follows the same formulation as Equation 4.1, but is applied locally within each window.

During inference, a strided sliding window strategy is adopted, in which each window contains N consecutive frames and overlaps the subsequent window by one frame. Let the first frame of window k define the local reference frame (frame 0) for that window. The pose of frame i within window k , obtained by recursively accumulating the predicted frame-to-frame transformations inside the window, is denoted as ${}_{(k)}^0T_i$.

To recover a continuous global trajectory, the locally accumulated poses must be related across consecutive windows. As an illustrative example, consider two consecutive windows, k and $k + 1$, which

share one overlapping frame. The accumulated pose of the last frame of window k is used to map the poses estimated in window $k + 1$ into the global coordinate system defined by window k , as formalized in Equation 4.3:

$${}^0T_i = {}^0T_{\ell_k} \cdot f_{k+1}T_i, \quad i = 1, \dots, N - 1, \quad (4.3)$$

where ℓ_k denotes the last frame of window k and f_{k+1} denotes the first frame of window $k + 1$. Since consecutive windows overlap by one frame, we have $f_{k+1} = \ell_k$. The same inter-window composition principle is then applied iteratively for all subsequent windows along the sequence, allowing all locally accumulated poses to be expressed in a common global reference frame and enabling reconstruction of the full trajectory.

During training, the GT labels are defined as frame-to-frame transformations and are computed in the same way as in the paired-frame configuration, following Equation 4.2.

Reference frame transformation

This output configuration is only applicable to the window-based input, as it requires a temporal window containing more than two frames to define a reference frame.

In this configuration, the network receives a window of N frames and predicts the pose of each frame relative to the first frame of that window. During training, the GT poses provided in world coordinate system are converted to the coordinate system of the reference frame (frame 0 of each window). This conversion from world coordinates to the reference frame follows the same formulation as in the previous case, with frame 0 used as the reference:

$${}^0T_i^{\text{GT}} = \left({}^wT_0^{\text{GT}}\right)^{-1} {}^wT_i^{\text{GT}}, \quad i = 1, \dots, N - 1, \quad (4.4)$$

where ${}^0T_i^{\text{GT}}$ denotes the GT transformation that maps points from the coordinate system of frame i into that of frame 0 of the window under analysis.

The inference procedure follows the same strided sliding window strategy described in the previous section for adjacent-frame transformations. Within each window k , the network predicts transformations relative to the first frame of that window (${}^0T_i$), which are then accumulated across overlapping windows into the global coordinate system of frame 0 of the sequence, as shown in Equation 4.3.

Multi-transformation output

Another approach explored in this work is inspired by the method proposed by Li *et al.* [58]. Instead of predicting a single transformation per frame within a N -frame window, this strategy introduces multiple transformation predictions within the window. Following the authors' formulation, the same set of predefined frame pairs is reproduced in this work, resulting in a total of 218 relative transformations for a 100-frame window. In this configuration, the model does not estimate a full set of poses with respect to all other frames; rather, it predicts only the transformations specified by these 218 selected pairs, which span a range of short and long-range relationships, including the final pair [98,99]. This subset forms a sparse set of motion constraints (a sparse graph), designed to capture temporal dependencies efficiently while remaining computationally feasible compared to estimating all possible frame combinations.

For the smaller window size ($N = 10$), the number of possible frame combinations is naturally much lower. To adapt the strategy to this reduced temporal context, two training variants are evaluated. In the first variant, all possible frame-to-frame combinations within the window are used (henceforth referred to as the full variant). In the second variant, a predefined subset of 26 pairwise combinations is selected (henceforth referred to as the sparse variant). This subset includes two global transformations (from the last frame to the first frame and from the middle frame to the first), all adjacent local transformations within the window, and additional sparse local transformations defined with different temporal gaps.

For inference with the 100-frame window, only one inference strategy is employed. Since the last prediction among the 218 estimated transformations corresponds to the pair [98, 99], the window is slid with a stride of 1, and only the predicted transformation $T_{98 \leftarrow 99}$ is retained at each step and accumulated to reconstruct the global trajectory (henceforth referred to as the *sparse_last* configuration). As a consequence, the first 98 frames of the US sequence are effectively discarded during reconstruction. Because each retained prediction corresponds to a transformation between consecutive frames (from frame i to frame $i - 1$), the global trajectory is obtained by recursively accumulating these predictions along the entire sequence, following the same formulation used in Equation 4.1.

For the 10-frame window, three different inference strategies are evaluated. Since both sparse and full training configurations include the adjacent frame pairs, this enables the use of frame-to-frame predictions (denoted as *sparse_previous* and *full_previous*, respectively). Although the network outputs all 26 (sparse variant) or 45 (full variant) transformations, only the 9 adjacent predictions are used, with the remaining ones discarded, as they are not required for this inference strategy. As in the adjacent frame variant, the GT labels are converted using Equation 4.2, and the predicted local transformations are accumulated within each window to obtain the pose of every frame relative to the first frame (Equation 4.1).

The second inference strategy follows the same principle as the 218-pairs approach used for the 100-frame window. As in that case, only the transformation between the last and the second-to-last frames of the window is retained (denoted as *sparse_last* and *full_last* for the sparse and full training variants, respectively). Since the window contains only 10 frames, this results in 8 discarded frames instead of the 98 discarded in the 100-frame configuration.

Finally, a third inference strategy is evaluated and is applied only to the full pair configuration. Unlike the sparse configuration, which does not provide predictions from every frame to the first frame of the window, the full set of 45 pairwise combinations allows such approach. In this strategy, hereafter referred to as the *full_reference* configuration, only the transformations mapping each frame directly to the first frame of the window are retained. This is equivalent to the reference frame output defined above, but it benefits from the additional temporal context provided by multiple transformation predictions. The global trajectory is then obtained using the update rule in Equation 4.3.

4.1.5 Ensembles

Ensemble methods have become increasingly popular in recent years, as combining multiple predictions, either from different models or from multiple outputs of the same model, often leads to more accurate and stable predictions [89].

In this dissertation, the potential benefits of ensembling are explored specifically in the context of pose prediction, evaluating whether different ensemble configurations could improve the accuracy and stability of pose predictions. To this end, two types of ensembling strategies were investigated: inference-level ensembles and model-level ensembles, as well as the combination of both approaches.

Inference-level ensemble

The first ensemble strategy explored in this thesis is inspired by an approach proposed by one of the TUS-REC challenge participants [85], here referred to as an *inference-level ensemble*.

Since the challenge dataset contains scans acquired in both scanning directions, a network trained on such data becomes capable of predicting the frame pose trajectory in either scanning direction. Taking advantage of this, the inference procedure is modified to include a bidirectional prediction step. First, the original frame sequence is passed through the network to obtain pose estimates for each frame with respect to a reference frame (frame 0). Then, the same set of frames is reversed in time and processed again by the same network. After obtaining the predictions from the reversed sequence, these poses are expressed with respect to the first frame of the reversed sequence, which corresponds to the last

frame of the original sequence. To make the forward and reversed predictions comparable, a change of reference frame is applied to map all reversed poses to the original reference frame. The forward and reversed estimates are then fused by computing the quaternion mean of the rotation components and the arithmetic mean of the translation vectors. The resulting per frame poses are accumulated using the same procedure as in the previous methods to obtain the final global trajectory.

Model-level ensemble

The second ensemble strategy explored in this work consists of combining multiple model checkpoints. Two configurations are evaluated: an ensemble formed from the four best performing models on the validation set, and an ensemble combining the best model with the last two checkpoints of the training run.

The four best models are selected based on the Euclidean distance computed on the validation data. For each validation scan, the Euclidean distance is computed point-wise between predicted and GT poses applied to a reference grid of points, whose definition is described in Section 4.2.2, and averaged over all points and all frames in the scan. The final validation distance is obtained by averaging this error across the validation dataset. This distance metric, expressed in millimeters, provides an interpretable measure of the mean reconstruction error and is therefore used to rank and select model checkpoints. Throughout training, the pool of the top four models is continuously updated by checking whether each newly saved checkpoint achieves a validation distance lower than the worst model currently in the set. When this occurs, the worst checkpoint is discarded and replaced by the new one. This selection process is performed independently of the fixed checkpoint saving interval, ensuring that the top performing models are retained regardless of when they are saved during training.

4.1.6 Loss functions

In this work, pose estimation is supervised by minimizing the discrepancy between predicted and GT point coordinates using the MSE loss applied to the full set of transformed reference image points. The MSE loss is widely adopted in the literature [58, 85] due to its simplicity and effectiveness in penalizing pointwise spatial errors, and it serves as the baseline loss function throughout this study.

However, relying exclusively on MSE may lead to overly smoothed motion estimates and can be insufficient to capture global consistency along a transformation sequence, particularly in the presence of accumulated prediction errors. To address these limitations, an alternative loss formulation combining MSE with a Pearson correlation term is also investigated. As reported in the literature [31, 55], the Pear-

son correlation term encourages stronger global correlation between predicted and GT trajectories, which can help mitigate drift effects in sequential pose estimation tasks.

Accordingly, two loss formulations are evaluated in this work: the standard MSE loss and a combined loss composed of MSE and a weighted Pearson correlation term. Their impact on pose prediction accuracy and trajectory stability is analyzed in the results and discussion section.

4.1.7 Implementations details

All proposed training strategies are implemented in PyTorch. A mini-batch gradient descent optimization scheme is used, with a batch size of 16 for the 10-frame window and adjacent frame input configurations, and a batch size of 8 for the 100-frame window (due to the higher memory requirements associated with the larger temporal context). Training is performed using the Adam optimizer, with a learning rate of 1×10^{-4} . All models are trained for 20000 epochs, following the same training protocol adopted in [58]. Checkpoints are saved every 100 epochs and additional checkpoints are retained according to the model selection strategies described earlier. Most experiments were executed on a Linux workstation equipped with an 18-core Intel Core i9-10980XE CPU, 125 GB of system memory, and a single NVIDIA Quadro RTX 6000 GPU with 24 GB of dedicated memory. The most computationally demanding configurations, namely those involving full resolution inputs and larger window sizes, were trained on an institutional high-performance computing server equipped with multiple NVIDIA A100 GPUs, providing a combined GPU memory pool of approximately 400 GB.

4.2 Results and discussion

4.2.1 Dataset

As mentioned above, the TUS-REC 2024 challenge dataset was used for the development and evaluation of the trackerless approaches, as mice data acquisition was still ongoing and the number of available scans was insufficient at the time to support model training and evaluation within the scope of this dissertation. This dataset contains US images acquired from the left and right forearms of 50 healthy volunteers. The images were collected using a BK Ultrasonix system (BK, Europe) with a curvilinear probe, an image resolution of 640×480 pixels, a frame rate of 20 fps, no speckle reduction, a gain of 48%, and an imaging depth of 9 cm.

For each subject, 24 scans were acquired following three predefined scanning trajectories: straight

line, C-shaped, and S-shaped, performed in both distal-to-proximal and proximal-to-distal directions, and with the probe oriented either perpendicular or parallel to the scanning direction.

Regarding the GT pose of each frame, the dataset provides accurate tracking obtained using an NDI Polaris Vicra optical tracking system (Northern Digital Inc., Canada), which records the 4×4 transformation matrix describing the pose of the probe for each frame in the camera coordinate system. The dataset also includes the calibration matrix defining the transformation between the US image coordinate system and the probe coordinate system [85].

To train the network, the dataset is split using a 3:1:1 ratio into training, validation, and testing subsets, following the strategy adopted by Li *et al.* [30]. This corresponds to 30 subjects for training, 10 for validation, and 10 for testing, resulting in 720 training scans, 240 validation scans, and 240 testing scans.

It is worth noting that the calibration procedure provided with the dataset, as well as the mathematical steps used to map the US image to the probe coordinate system, follow the same underlying principles as those employed in this work for the Aurora MTS during mice acquisitions (Section 3.1.2).

4.2.2 Evaluation metrics

For each of the methodologies evaluated in this work, a set of six metrics was recorded to assess pose estimation accuracy and its impact on reconstruction quality. Four of these metrics — global pixel reconstruction error (GPE), global corner reconstruction error (GCE), local pixel reconstruction error (LPE), and local corner reconstruction error (LCE), followed the definitions provided in the challenge paper [85]. In addition to these, two complementary measures were included: the FD and the HD95, as described in Section 2.4.1, with HD95 computed at the global level to better capture long term drift and near worst case deviations in the predicted trajectories.

Before describing each evaluation metric, it is important to clarify how global and local predictions were interpreted and computed, and why both representations are required. Evaluating global metrics is essential to analyze drift across the entire sequence, while local metrics assess the reconstruction quality at the frame level. Global predictions represent the pose of each frame with respect to a reference frame (frame 0). As explained earlier, if a model outputs only local transformations, or if it predicts global poses only within a sliding window, these predictions must be accumulated to obtain a consistent global trajectory for the entire sequence.

Local predictions, on the other hand, represent the transformation between consecutive frames. For methods that output global poses (*i.e.*, the pose of each frame with respect to frame 0), local frame-to-frame transformations are obtained by computing the relative transformation between successive global

poses. Specifically, the local transformation between frames i and $i + 1$ is computed as:

$${}^i T_{i+1}^{\text{local}} = (T_i^{\text{global}})^{-1} T_{i+1}^{\text{global}}, \quad i = 0, \dots, N - 2. \quad (4.5)$$

where T_i^{global} and T_{i+1}^{global} denote the global poses of frames i and $i + 1$, respectively, which could correspond either to predicted or GT poses.

The final global and local vectors (both predicted and GT) used to compute the evaluation metrics are compared in the probe coordinate system. To this end, the predicted and GT transformations are applied to a set of reference image points, which are first converted from pixel coordinates to physical units and mapped from the image coordinate system to the probe coordinate system.

This conversion is performed using a calibration matrix that combines a rigid transformation, which maps image coordinates to the probe coordinate system (T_{rigid}), and a scale transformation, which converts pixel units into millimeters (T_{scale}), as defined in Equation 4.6:

$$T_{\text{calib}} = T_{\text{rigid}} T_{\text{scale}}. \quad (4.6)$$

Depending on the metric, the reference point set consist either of all points in the image grid or only of the four corner points, as detailed in the corresponding metric definitions below.

Using this calibration transformation, the reconstructed 3D points are obtained by multiplying the predicted / GT pose with the calibration matrix and the set of reference image points, where $U = \{u_1, u_2, \dots, u_M\}$ is the set of homogeneous image points, with $u_i = [u_i, v_i, 0, 1]^T$. The transformed 3D points are then obtained as:

$$P_{\text{pred/GT}} = T_{\text{pred/GT}} T_{\text{calib}} U. \quad (4.7)$$

Now that all preprocessing steps have been described, the computation of each evaluation metric can be detailed. The first two metrics, the GPE and the GCE, quantify the global performance of the reconstruction. The difference between them lay in the set of points used to compute the error: the GPE is computed over the entire image grid, whereas the GCE is evaluated only in the four corner points of the frame. Assessing the global error on these two distinct sets is important because corner points are located farther from the image center and therefore tend to exhibit larger deviations, providing a more pessimistic and informative estimate of the maximum expected reconstruction error.

For both metrics, the reconstruction error was computed for each test acquisition separately. Let $P_{s,i}^{\text{pred}}(u)$ and $P_{s,i}^{\text{GT}}(u)$ denote, respectively, the predicted and GT 3D coordinates of the point corresponding to image location u in frame i of scan s , with $u \in U$. The reconstruction error for scan s is thus

defined as:

$$e_s = \frac{1}{N_s M_s} \sum_{i=1}^{N_s} \sum_{u \in U_s} \|P_{s,i}^{\text{global_pred}}(u) - P_{s,i}^{\text{global_GT}}(u)\|_2, \quad (4.8)$$

where N_s is the number of reconstructed frames in scan s , and M_s is the number of points used for the metric (all pixels for GPE, or the four corners for GCE). The final GPE and GCE values reported in this work correspond to the mean of e_s across all test acquisitions, with the associated standard deviation computed over the same set of scans.

Similar to these global metrics, the LPE and LCE were computed using the same formulation as in Equation 4.8, but replacing the global reconstructed points by the corresponding local reconstructed points. In this case, the error quantifies the frame-to-frame reconstruction accuracy rather than the drift accumulated over the entire sequence.

Regarding the FD error, this metric quantifies the accumulated error at the end of the sweep and is defined as the Euclidean distance between the predicted and GT positions of the four corner points of the last frame (Equation 4.9). In practical terms, FD measures how much frame-to-frame errors compound over time, providing a direct indicator of long-term drift in the reconstruction.

$$\text{FD} = \|\bar{t}_{N-1}^{\text{pred}} - \bar{t}_{N-1}^{\text{GT}}\|. \quad (4.9)$$

The reported value corresponds to the mean (and standard deviation) computed over all test acquisitions.

Finally, the HD95 was computed to evaluate the near worst case between the predicted and GT trajectories across the entire sweep (Equation 2.10). Unlike global error metrics that reflect average accuracy, the HD95 corresponds to the 95th percentile of the distances between the two trajectories, thereby mitigating the impact of extreme outliers while still capturing large discrepancies. Given two point sets A (predicted trajectory) and B (GT trajectory), the symmetric HD95 is defined as:

$$\text{HD95}(A, B) = \max \left\{ P_{95} \left(\min_{b \in B} d(a, b) \mid a \in A \right), P_{95} \left(\min_{a \in A} d(b, a) \mid b \in B \right) \right\}. \quad (4.10)$$

4.2.3 Network performance

In this section, an overview of the network performance is provided, comparing the different input and output configurations and discussing their impact on the predicted poses. The results obtained from the inference stage are summarized in Table 5. For a fair comparison across all methods, all inferences were

Table 5: Performance comparison for different input and output strategies.

Input	Output	Res.	GPE	GCE	LPE	LCE	FD	HD95
2	—	Orig.	6.305 ± 2.858	7.236 ± 2.980	0.135 ± 0.020	0.153 ± 0.153	13.922 ± 5.844	13.370 ± 6.253
		res./4	59.529 ± 12.966	59.917 ± 12.867	0.378 ± 0.054	0.390 ± 0.053	104.530 ± 19.721	104.647 ± 19.208
10	previous	Orig.	5.241 ± 3.277	6.212 ± 3.312	0.119 ± 0.020	0.136 ± 0.021	11.566 ± 5.680	12.425 ± 7.196
		Res./4	61.320 ± 12.030	61.681 ± 12.021	0.387 ± 0.047	0.397 ± 0.047	107.251 ± 21.597	107.349 ± 18.619
	reference	Orig.	4.743 ± 2.782	5.617 ± 2.879	0.122 ± 0.018	0.140 ± 0.020	11.010 ± 5.515	11.079 ± 6.745
		Res./4	6.096 ± 3.022	7.007 ± 3.106	0.136 ± 0.019	0.154 ± 0.021	13.135 ± 5.687	13.444 ± 6.967
	sparse_previous	Orig.	5.627 ± 3.127	6.607 ± 3.221	0.118 ± 0.019	0.135 ± 0.020	13.506 ± 6.089	13.158 ± 6.963
		Res./4	8.936 ± 4.196	9.927 ± 4.177	0.146 ± 0.022	0.165 ± 0.023	19.056 ± 7.284	19.538 ± 8.715
	sparse_last	Orig.	7.757 ± 3.861	8.473 ± 3.866	0.126 ± 0.020	0.144 ± 0.021	16.537 ± 7.413	16.405 ± 8.594
		Res./4	11.311 ± 5.711	12.205 ± 5.536	0.154 ± 0.023	0.171 ± 0.024	24.228 ± 11.118	24.410 ± 11.514
	full_previous	Orig.	4.981 ± 3.340	5.732 ± 3.357	0.119 ± 0.019	0.136 ± 0.020	11.128 ± 6.086	11.521 ± 7.236
		Res./4	7.376 ± 3.849	8.237 ± 3.893	0.140 ± 0.021	0.158 ± 0.023	15.608 ± 6.824	16.074 ± 8.994
	full_last	Orig.	7.489 ± 3.885	8.008 ± 3.846	0.128 ± 0.019	0.145 ± 0.021	14.974 ± 6.747	15.336 ± 8.455
		Res./4	11.371 ± 5.146	12.261 ± 5.303	0.153 ± 0.023	0.170 ± 0.025	24.801 ± 11.636	23.212 ± 11.935
	full_reference	Orig.	4.934 ± 3.274	5.679 ± 3.289	0.118 ± 0.019	0.135 ± 0.020	11.009 ± 5.950	11.389 ± 7.147
		Res./4	6.915 ± 3.860	7.701 ± 3.866	0.137 ± 0.022	0.154 ± 0.023	14.654 ± 6.557	15.159 ± 8.676
100	previous	Orig.	7.934 ± 3.387	8.769 ± 3.402	0.161 ± 0.022	0.178 ± 0.024	16.043 ± 6.594	15.309 ± 6.700
		Res./4	15.879 ± 8.618	16.640 ± 8.525	0.201 ± 0.045	0.218 ± 0.046	29.900 ± 15.956	28.234 ± 14.792
	reference	Orig.	6.810 ± 2.911	7.650 ± 2.998	0.161 ± 0.020	0.142 ± 0.022	14.040 ± 5.476	13.817 ± 6.565
		Res./4	10.757 ± 6.045	11.601 ± 6.075	0.178 ± 0.032	0.195 ± 0.033	21.418 ± 11.024	19.916 ± 10.008
	sparse_last	Orig.	7.757 ± 3.861	8.473 ± 3.865	0.126 ± 0.019	0.144 ± 0.021	16.537 ± 7.413	16.405 ± 8.594
		Res./4	18.748 ± 7.922	20.199 ± 7.764	0.200 ± 0.037	0.219 ± 0.038	37.771 ± 14.243	34.023 ± 12.657

Values are reported as mean ± standard deviation. Orig. refers to the dataset at its original spatial resolution, whereas Res./4 corresponds to a version downsampled by a factor of four. Bold values indicate the lowest error achieved for each metric across all evaluated configurations..

evaluated exclusively on the frames that were common to every approach. All results were obtained using models trained with the MSE loss.

A first observation was that, across all methods, training the network using the original resolution consistently led to better performance. This outcome was expected as downsampling the acquired frames inevitably removes important visual details, including fine-grained speckle patterns. Since speckle information has been shown in the literature to provide useful cues for learning spatial relationships and improving pose estimation, preserving the native resolution allowed the models to better capture these patterns and, consequently, to produce more accurate predictions. This finding is consistent with the results reported by Prevost *et al.* [45], who demonstrated that speckle patterns contribute positively to the performance and generalization ability of CNN for US motion estimation. In their study, networks trained on unfiltered images containing speckle noise outperformed those trained on filtered data. These results reinforce the

importance of preserving native image resolution and speckle information when learning spatial transformations from US data.

It is also important to note that, in some cases, the performance difference between downsampled and original resolutions was very substantial. In particular, in certain configurations the network appeared unable to learn how the frames should align with one another. This behavior was observed for the paired-frame input, where only two consecutive frames are provided, as well as for the 10-frame window when predicting poses relative to the previous frame. This effect could be explained by the combination of two factors: a strong dependence on image resolution and the limited temporal context available in these approaches. In the first case, where only two frames were provided at a time, the model cannot leverage any past or future motion, and the two input frames were often highly similar due to their temporal proximity. As a result, the network received insufficient temporal variation to produce a reliable estimation of the relative motion, a limitation that became even more pronounced when the input resolution was reduced.

In the second case, although the model processes a window of 10 frames, it is still required to learn frame-to-frame relationships by predicting each pose relative to its immediate predecessor. When the images are downsampled, this learning task becomes significantly more challenging, again highlighting the importance of both high spatial resolution and sufficient temporal context for accurate pose estimation. This interpretation is further supported by the results obtained with the 100-frame window approach, which also relies on predictions relative to the previous frame. Although the full-resolution configuration achieves the best overall performance, the downsampled $\times 4$ version remains reasonably accurate. This suggests that a larger temporal context provides complementary information that helps compensate for the loss of spatial detail introduced by downsampling.

Another aspect influencing inference performance is the choice of reference frame used for pose prediction. As described earlier, the network can estimate the pose of each frame either relative to a fixed reference frame or relative to the previous frame within a 10- or 100-frame window. It was observed that both GPE and GCE achieved lower values when poses were predicted with respect to a fixed reference frame, compared to predictions made relative to the previous frame. This behavior is expected, as predicting all poses relative to a single reference inherently limits drift accumulation, a conclusion that is further supported by the lower FD and HD95 values obtained for this configuration.

Interestingly, the configuration with $N = 10$ achieved the lowest errors overall, suggesting that while temporal context is beneficial, excessively large windows may introduce redundancy or increased variability that complicates optimization. A clear trade-off was observed, with the 100-frame window outperforming the paired-frame input but underperforming the 10-frame window. One possible explanation lies in the lim-

ited capacity of the early layers of *EfficientNet-B1*, whose first convolutional layer contains only 32 filters. When too many frames are provided simultaneously, the diversity and variability of visual patterns may exceed the expressive capacity of these early layers, making feature extraction and optimization more difficult. In contrast, the 10-frame window appears to provide sufficient temporal context without overwhelming the network, resulting in the best overall performance.

As discussed in the previous paragraph, the approaches that achieved the best LPE and LCE values were those using the 10-frame window. In particular, the three best-performing configurations correspond to the multi-transformation strategies: *sparse_previous* and the full variants (*full_previous* and *full_reference*). These results are consistent with earlier observations: although the *full_reference* approach uses only global predictions during inference, the network is trained to learn all pairwise relationships within the window. This richer supervision not only strengthens the model’s ability to capture fine-scale motion, but also implicitly enforces geometric consistency across transformations, as the model learns that longer-range motions should be consistent with the composition of multiple shorter-range transformations within the window. As a result, this formulation helps mitigate drift accumulation, ultimately leading to improved local accuracy compared with methods relying solely on independent global predictions.

When comparing the perpendicular and parallel scanning directions, a clear difference in performance was observed. Reconstructions obtained from perpendicular scans were consistently more accurate and stable, whereas parallel scans showed larger deviations and more drift. This difference arises because perpendicular motion produces stronger changes in speckle patterns and anatomical structures, providing the network with richer visual cues to estimate the probe’s movement and making motion estimation easier and more reliable. In contrast, when the probe moves parallel to the arm, the frames become much more similar to each other. The same structures appear in nearly the same position for a long time, and the speckle pattern changes only minimally. As a result, the network receives fewer distinctive features to guide pose estimation, making the task far more ambiguous and ultimately more prone to drift.

To better illustrate these observations, representative examples were selected from a single subject (patient 0) using the best-performing configuration. Based on the global error metrics, which are considered the most relevant in this work because they reflect the accumulated error over the full scan, the 10-frame window approach predicting poses relative to a reference frame was identified as the best-performing model and was therefore selected for qualitative analysis. Figures 40 and 41 present, respectively, the best and worst reconstruction examples for each trajectory type (linear, S-shaped, and C-shaped). These qualitative results confirm the previously described behavior: all worst cases correspond to parallel scanning trajec-

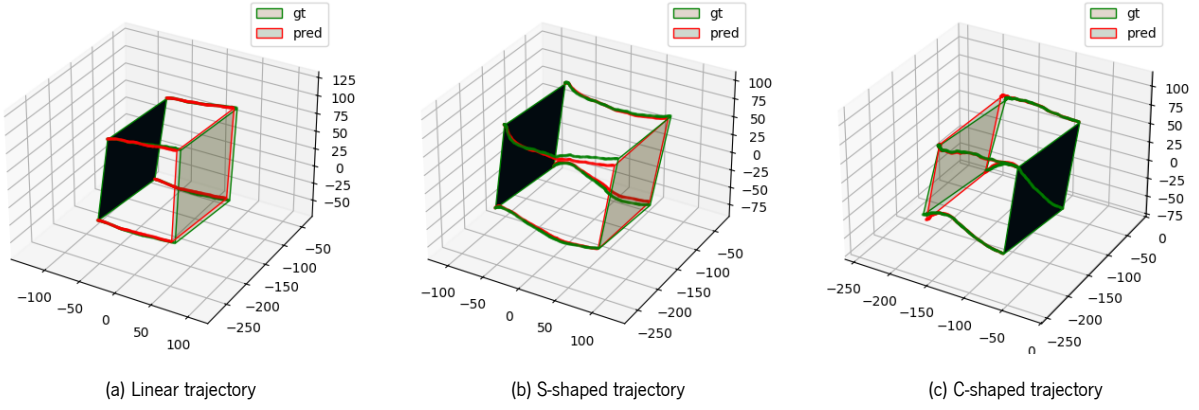


Figure 40: Best three inference cases for patient 0 obtained with the 10-frame window model predicting poses to a reference frame (original resolution), covering linear, S-shaped and C-shaped trajectories.

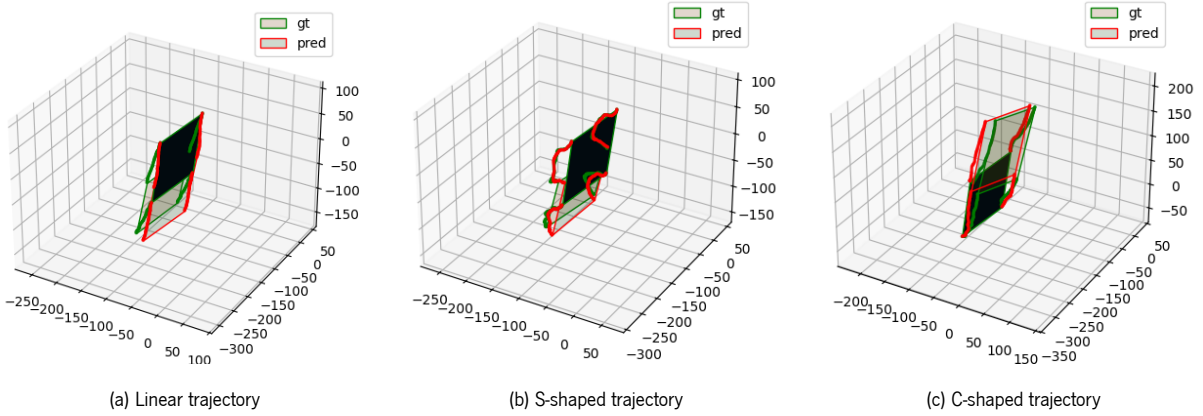


Figure 41: Worst three inference cases for patient 0 obtained with the 10-frame window model predicting poses to a reference frame (original resolution), covering linear, S-shaped and C-shaped trajectories.

ries, where clear deviations and drift are observed, whereas perpendicular scans yield substantially more accurate and stable reconstructions.

Once the best model was identified, ensemble techniques were applied exclusively to this configuration to assess whether additional performance gains could be achieved, as summarized in Table 6.

Notably, among the evaluated ensemble strategies, the bidirectional ensemble produced one of the strongest overall results, achieving the best performance in the GCE, FD and HD95 metrics, as well as the second lowest GPE. These metrics reflect a substantial reduction in drift and indicate that even the worst-case deviations between the predicted and GT trajectories (as captured by HD95) were considerably reduced. This highlights the benefit of incorporating reversed frame predictions, which provide complementary information and help mitigate directional bias in pose estimates.

Among all configurations, the combination of the bidirectional ensemble with the “best checkpoint + two latest models” achieved the best overall global performance, obtaining the lowest GPE and the

Table 6: Performance of all ensemble configurations for the best-performing model, trained with original resolution images using a 10-frame window.

Model ensemble	Inference ensemble	GPE	GCE	LPE	LCE	FD	HD
–	Unidirectional	4.743 $\pm$ 2.782	5.618 $\pm$ 2.879	0.122 $\pm$ 0.018	0.140 $\pm$ 0.020	11.010 $\pm$ 5.515	11.079 $\pm$ 6.745
–	Bidirectional	4.599 $\pm$ 2.787	5.405 $\pm$ 2.844	0.123 $\pm$ 0.020	0.140 $\pm$ 0.021	10.607 $\pm$ 5.169	10.959 $\pm$ 6.481
Top-4 models	Unidirectional	4.809 $\pm$ 2.737	5.667 $\pm$ 2.836	0.120 $\pm$ 0.018	0.137 $\pm$ 0.019	11.031 $\pm$ 5.275	11.338 $\pm$ 6.549
Best + 2 last models	Unidirectional	4.659 $\pm$ 2.638	5.547 $\pm$ 2.752	0.119 $\pm$ 0.018	0.136 $\pm$ 0.020	10.779 $\pm$ 5.173	11.029 $\pm$ 6.447
Top-4 models	Bidirectional	4.748 $\pm$ 3.019	5.535 $\pm$ 3.066	0.122 $\pm$ 0.020	0.139 $\pm$ 0.021	10.906 $\pm$ 5.367	11.359 $\pm$ 6.873
Best + 2 last models	Bidirectional	4.586 $\pm$ 2.897	5.408 $\pm$ 2.961	0.122 $\pm$ 0.020	0.139 $\pm$ 0.021	10.619 $\pm$ 5.251	11.047 $\pm$ 6.576

Values are reported as mean $\pm$ standard deviation. Bold values indicate the lowest error achieved for each metric across all evaluated configurations.

second lowest GCE and FD values. This suggests that merging forward and reversed predictions with model checkpoints saved at different stages of training helps to further stabilize motion estimation by compensating for variability that individual models tend to propagate.

In contrast, the ensemble formed by the best-performing checkpoint together with the two latest saved models achieved the lowest local errors (LPE and LCE). However, it is important to note that the differences observed across ensemble variants for these local metrics are very small and likely not statistically significant. This suggests that, at the frame-to-frame level, all ensemble configurations perform similarly in terms of local motion estimation. The benefits of the ensemble strategies therefore do not primarily manifest in the local metrics themselves, but rather in how these small local errors accumulate over time. As discussed above, the improvements introduced by the ensemble approaches become more evident in the global metrics, where reduced drift and improved long-term consistency can be observed.

In parallel with the ensemble study, the influence of the training loss formulation on model performance was also investigated. To this end, two representative models were selected, both using a 10-frame window input but differing in their output strategies: one predicting poses relative to the previous frame (*previous output configuration*) and the other predicting poses relative to a fixed reference frame (*reference output configuration*). These configurations were chosen because they exhibited competitive performance while relying on fundamentally different pose prediction strategies. For each model, multiple loss configurations combining MSE with different Pearson correlation weights were evaluated, as summarized in Table 7.

For the previous configuration, where poses are predicted locally between consecutive frames, drift accumulation is more pronounced due to the recursive integration of frame-to-frame transformations. In this setting, incorporating a Pearson correlation term proved beneficial, leading to a clear reduction in drift-related metrics such as FD and HD95. This suggests that the correlation term effectively promotes global trajectory coherence, compensating for errors that accumulate over time in local prediction schemes.

Table 7: Performance comparison for different Pearson correlation weights combined with MSE for two representative 10-frame window models (previous and reference outputs) trained at original resolution.

Model	Pearson weight	GPE	GCE	LPE	LCE	FD	HD95
Previous	–	5.241 $\pm$ 3.277	6.212 $\pm$ 3.312	0.119 $\pm$ 0.020	0.136 $\pm$ 0.021	11.566 $\pm$ 5.680	12.425 $\pm$ 7.196
	4000	5.584 $\pm$ 3.111	6.557 $\pm$ 3.162	0.121 $\pm$ 0.020	0.139 $\pm$ 0.022	12.675 $\pm$ 5.516	13.593 $\pm$ 7.437
	12000	4.364 $\pm$ 2.560	5.226 $\pm$ 2.658	0.117 $\pm$ 0.020	0.134 $\pm$ 0.021	10.180 $\pm$ 5.122	10.253 $\pm$ 6.017
	20000	6.152 $\pm$ 3.477	6.933 $\pm$ 3.493	0.123 $\pm$ 0.020	0.140 $\pm$ 0.021	13.365 $\pm$ 6.128	13.312 $\pm$ 7.195
Reference	–	4.743 $\pm$ 2.782	5.617 $\pm$ 2.879	0.122 $\pm$ 0.018	0.140 $\pm$ 0.020	11.010 $\pm$ 5.515	11.079 $\pm$ 6.745
	4000	5.222 $\pm$ 3.655	6.097 $\pm$ 3.698	0.126 $\pm$ 0.020	0.144 $\pm$ 0.021	11.034 $\pm$ 5.476	11.391 $\pm$ 6.731
	12000	5.529 $\pm$ 3.086	6.309 $\pm$ 3.085	0.126 $\pm$ 0.018	0.143 $\pm$ 0.019	11.742 $\pm$ 6.700	12.067 $\pm$ 7.711
	20000	5.334 $\pm$ 2.801	6.254 $\pm$ 2.875	0.125 $\pm$ 0.019	0.142 $\pm$ 0.020	12.340 $\pm$ 5.741	12.246 $\pm$ 6.412

Values are reported as mean $\pm$ standard deviation. Bold values indicate the lowest error achieved for each metric across all evaluated configurations.

In contrast, for the reference configuration, poses are predicted directly with respect to a fixed reference frame, inherently limiting drift accumulation. As a result, the standard MSE loss already provides stable and accurate supervision, and the addition of a Pearson correlation term does not yield further improvements.

4.2.4 Conclusion

The results obtained in this chapter demonstrate that both spatial resolution and temporal context are fundamental to achieve accurate trackerless pose estimation. Models trained using the original image resolution consistently outperformed those trained on downsampled images, and a 10-frame window provided the best balance between temporal context and model capacity. Predicting all poses relative to a fixed reference frame (frame 0 in this work) also helped reduce drift, as it prevents error accumulation over time. Alternatively, poses can be predicted relative to the preceding frame, which enables a purely local formulation of the motion estimation problem but is more susceptible to drift due to the recursive accumulation of frame-to-frame errors. In this case, the training loss formulation becomes particularly important, and the inclusion of a Pearson correlation term was shown to help mitigate drift by encouraging greater global consistency along the predicted trajectory.

Finally, ensemble strategies offered small but consistent improvements, with bidirectional prediction enhancing global stability and checkpoint diversity improving local accuracy. Overall, the 10-frame reference model at full resolution, combined with ensemble refinement, emerged as the most robust method in this study. Although the approach could not yet be evaluated on the mice dataset, the results obtained on the challenge dataset indicate that the proposed trackerless framework is promising and may be well suited for future application in small animal freehand imaging.

Chapter 5

Conclusion and future developments

This thesis presented a complete framework for freehand 3D US reconstruction exploring both tracker-based and trackerless methodologies, with the aim of enabling reliable volumetric imaging in small animal studies. The work began by establishing a validated acquisition pipeline using the Aurora electromagnetic MTS, including a custom-designed validation phantom, detailed calibration procedures, and a rigorous assessment of spatial and temporal accuracy. The results obtained with this tracker-based setup demonstrate that, when properly calibrated, electromagnetic tracking provides sufficiently accurate and stable pose estimates to support high-quality volumetric reconstruction in controlled experimental conditions. This validated pipeline therefore serves both as a reliable reconstruction solution and as a strong reference for evaluating trackerless approaches.

In parallel, several DL approaches for trackerless pose estimation were developed and systematically evaluated. Using the TUS-REC challenge dataset for model development, the study explored a wide range of configurations, including paired-frame and window-based input strategies, as well as multi-transformation prediction schemes. The results demonstrated that two factors are particularly critical for accurate pose estimation: preserving the original image resolution and providing an appropriate amount of temporal context. Among the evaluated configurations, a model using a 10-frame window input and predicting poses relative to a reference frame, combined with model- and inference-level ensemble strategies, achieved strong overall performance, offering a good balance between global stability, local accuracy, and computational feasibility. However, the best overall performance was achieved by the 10-frame window model predicting poses relative to the previous frame when a combined MSE loss with a weighted Pearson correlation term was employed. In this configuration, the Pearson loss significantly improved global consistency and reduced drift, allowing the locally predicted (*previous*) model to outperform all other configurations across all evaluated metrics.

The ensemble experiments further showed that the combination of predictions from multiple model checkpoints or scanning directions (bidirectional) leads to consistent, albeit moderate, performance gains.

In particular, bidirectional inference proved effective in reducing drift and improving global trajectory coherence, while checkpoint diversity contributed to stabilizing frame-to-frame motion estimation. These findings highlight the importance of leveraging complementary information to mitigate error accumulation in trackerless US pose estimation.

In summary, this work presents a comprehensive study of both hardware-based and DL-based pose estimation for freehand US reconstruction, identifying the key factors that influence reconstruction quality and the model configuration that best balances accuracy.

Despite these promising results, several limitations should be acknowledged. Due to time constraints and the limited availability of mice acquisitions, the trackerless approaches were not evaluated on the mice dataset within the scope of this dissertation. Moreover, the experiments revealed a consistent performance gap between perpendicular and parallel scanning trajectories, with substantially higher errors observed in parallel sweeps. This behavior is likely related to the reduced variation in anatomical structures and speckle patterns during parallel motion, which makes pose estimation more ambiguous. Given that the ultimate goal of this work is to estimate probe motion for the reconstruction of subcutaneous tumors that may lack consistent shapes or well defined boundaries, this limitation could become particularly relevant in small animal imaging. These observations suggest that the current models may partially rely on learning appearance-based cues rather than exclusively modeling probe motion.

Future work should therefore focus on addressing these limitations. A critical next step is the evaluation of the proposed trackerless models on the mice dataset once sufficient data become available. In addition, extending the evaluation to in vivo tumor acquisitions is essential to assess performance under realistic imaging conditions. Further improvements may be achieved through more advanced data augmentation strategies, alternative network architectures with stronger temporal modeling capabilities, or the design of loss functions that explicitly enforce motion consistency. Together, these developments would contribute to improving robustness and generalization, bringing trackerless freehand US reconstruction closer to practical deployment in preclinical imaging studies.

Bibliography

- [1] Roderic Guigó Alessandra Breschi, Thomas R. Gingeras. Comparative transcriptomics in human and mouse. *Nature Reviews Genetics*, 18(7):425–440, 2017.
- [2] John G Vandenberg. Use of house mice in biomedical research. *ILAR Journal*, 41(3):133–135, 2000.
- [3] Adriana Domínguez-Oliva, Ismael Hernández-Avalos, Julio Martínez-Burnes, Adriana Olmos-Hernández, Antonio Verduzco-Mendoza, and Daniel Mota-Rojas. The importance of animal models in biomedical research: current insights and applications. *Animals*, 13(7):1223, 2023.
- [4] World Health Organization. Global cancer burden growing amidst mounting need for services, 2024. Available at: <https://www.who.int/news/item/01-02-2024-global-cancer-burden-growing-amidst-mounting-need-for-services>. [Accessed: 07-October-2024].
- [5] Jeffrey R Ashton, Jennifer L West, and Cristian T Badea. In vivo small animal micro-CT using nanoparticle contrast agents. *Frontiers in Pharmacology*, 6:256, 2015.
- [6] Judith E Spiro, Miriam Rinneburger, Dennis M Hedderich, Mladen Jokic, Hans Christian Reinhardt, David Maintz, Moritz Palmowski, and Thorsten Persigehl. Monitoring treatment effects in lung cancer-bearing mice: clinical CT and clinical MRI compared to micro-CT. *European Radiology Experimental*, 4:1–8, 2020.
- [7] Navid Farr, Yak-Nam Wang, Samantha D’Andrea, Kayla M Gravelle, Joo Ha Hwang, and Donghoon Lee. Noninvasive characterization of pancreatic tumor mouse models using magnetic resonance imaging. *Cancer medicine*, 6(5):1082–1090, 2017.
- [8] Xiangzhi Zhou, Haihua Bao, Ruqayyah Al-Hashem, Hongbin Ji, Mitchell Albert, Kwok-Kin Wong, and Yanping Sun. Magnetic resonance imaging of the response of a mouse model of non-small cell lung cancer to tyrosine kinase inhibitor treatment. *Comparative Medicine*, 58(3):276–281, 2008.

- [9] Sarah E Shelton, Jodi Stone, Fei Gao, Donglin Zeng, and Paul A Dayton. Microvascular ultrasonic imaging of angiogenesis identifies tumors in a murine spontaneous breast cancer model. *International Journal of Biomedical Imaging*, 2020(1):7862089, 2020.
- [10] Jie Ni, Paul Cozzi, Tzong-Tyng Hung, Jingli Hao, Peter Graham, and Yong Li. Monitoring prostate tumor growth in an orthotopic mouse model using three-dimensional ultrasound imaging technique. *Translational Oncology*, 9(1):41–45, 2016.
- [11] Sharad S Singhal, Rachana Garg, Atish Mohanty, Pankaj Garg, Sravani Keerthi Ramisetty, Tamara Mirzapioazova, Raffaella Soldi, Sunil Sharma, Prakash Kulkarni, and Ravi Salgia. Recent advancement in breast cancer research: insights from model organisms—mouse models to zebrafish. *Cancers*, 15(11):2961, 2023.
- [12] Cristian T Badea. *Principles of Micro X-ray computed tomography*, pages 47–64. Elsevier, 2021.
- [13] Joel D Boerckel, Devon E Mason, Anna M McDermott, and Eben Alsberg. Microcomputed tomography: approaches and applications in bioengineering. *Stem Cell Research & Therapy*, 5:1–12, 2014.
- [14] Lyubomir Zagorchev, Pierre Oses, Zhen W Zhuang, Karen Moodie, Mary Jo Mulligan-Kehoe, Michael Simons, and Thierry Couffignal. Micro computed tomography for vascular exploration. *Journal of Angiogenesis Research*, 2:1–11, 2010.
- [15] Federico Bruno, Vincenza Granata, Flavia Cobiauchi Bellisari, Ferruccio Sgalambro, Emanuele Tomasino, Pierpaolo Palumbo, Francesco Arrigoni, Diletta Cozzi, Francesca Grassi, Maria Chiara Brunese, et al. Advanced magnetic resonance imaging techniques: technical principles and applications in nanomedicine. *Cancers*, 14(7):1626, 2022.
- [16] Syifa Azhar and Le Roy Chong. Clinician’s guide to the basic principles of MRI. *Postgraduate Medical Journal*, 99(1174):894–903, 2023.
- [17] Wingchi E Kwok. Basic principles of and practical guide to clinical MRI radiofrequency coils. *RadioGraphics*, 42(3):898–918, 2022.
- [18] Patrick McConville. Small animal preparation and handling in MRI. *In vivo NMR imaging: methods and protocols*, pages 89–113, 2011.

- [19] Marc A Brockmann, Andre Kemmling, and Christoph Groden. Current issues and perspectives in small rodent magnetic resonance imaging using clinical MRI scanners. *Methods*, 43(1):79–87, 2007.
- [20] Pak-Hei Yeung, Linde S. Hesse, Moska Aliasi, Monique C. Haak, Weidi Xie, and Ana I.L. Namburete. Sensorless volumetric reconstruction of fetal brain freehand ultrasound scans with deep implicit representation. *Medical Image Analysis*, 94:103147, 2024.
- [21] Tiexiang Wen, Qingsong Zhu, Wenjian Qin, Ling Li, Fan Yang, Yaoqin Xie, and Jia Gu. An accurate and effective FMM-based approach for freehand 3d ultrasound reconstruction. *Biomedical Signal Processing and Control*, 8(6):645–656, 2013.
- [22] Susannah J Patey and James P Corcoran. Physics of ultrasound. *Anaesthesia & Intensive Care Medicine*, 22(1):58–63, 2021.
- [23] Katerina Stripling, Finn Timmermans, Sofiane Décombas-Deschamps, Jérémy H Thalgott, David Lemonnier, Margreet R de Vries, Ton J Rabelink, Mickael Tanter, Thomas Deffieux, and Franck Lebrin. Imaging and quantification of the hepatic vasculature of mice using ultrafast doppler ultrasound. *Jove*, 2024.
- [24] Qinghua Huang, Yongping Zheng, Minhua Lu, Tianfu Wang, and Siping Chen. A new adaptive interpolation algorithm for 3D ultrasound imaging with speckle reduction and edge preservation. *Computerized Medical Imaging and Graphics*, 33(2):100–110, 2009.
- [25] Xiankang Chen, Tiexiang Wen, Xingmin Li, Wenjian Qin, Donglai Lan, Weizhou Pan, and Jia Gu. Reconstruction of freehand 3D ultrasound based on kernel regression. *BioMedical Engineering Online*, 13:1–15, 2014.
- [26] Mohammad I Daoud, Abdel-Latif Alshalalfah, Falah Awwad, and Mahasen Al-Najar. Freehand 3D ultrasound imaging system using electromagnetic tracking. In *2015 International Conference on Open Source Software Computing (OSSCOM)*, pages 1–5. IEEE, 2015.
- [27] Mohammad Hamed Mozaffari and Won-Sook Lee. Freehand 3D ultrasound imaging: a systematic review. *Ultrasound in Medicine & biology*, 43(10):2099–2124, 2017.
- [28] Yifan Wang, Tianyu Fu, Chan Wu, Jingfan Fan, Hong Song, Deqiang Xiao, Yucong Lin, Fangyi Liu, and Jian Yang. Adaptive tetrahedral interpolation for reconstruction of uneven freehand 3D ultrasound. *Physics in Medicine & Biology*, 68(5):055005, 2023.

- [29] Raphael Prevost, Mehrdad Salehi, Simon Jagoda, Navneet Kumar, Julian Sprung, Alexander Ladikos, Robert Bauer, Oliver Zettinig, and Wolfgang Wein. 3D freehand ultrasound without external tracking using deep learning. *Medical Image Analysis*, 48:187–202, 2018.
- [30] Qi Li, Ziyi Shen, Qianye Yang, Dean C Barratt, Matthew J Clarkson, Tom Vercauteren, and Yipeng Hu. Nonrigid reconstruction of freehand ultrasound without a tracker. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 689–699. Springer, 2024.
- [31] Mingyuan Luo, Xin Yang, Hongzhang Wang, Haoran Dou, Xindi Hu, Yuhao Huang, Nishant Ravikumar, Songcheng Xu, Yuanji Zhang, Yi Xiong, et al. Recon: Online learning for sensorless freehand 3D ultrasound reconstruction. *Medical Image Analysis*, 87:102810, 2023.
- [32] Jiahui Dong, Tianyu Fu, Yucong Lin, Qiaoling Deng, Jingfan Fan, Hong Song, Zhigang Cheng, Ping Liang, Yongtian Wang, and Jian Yang. Hole-filling based on content loss indexed 3D partial convolution network for freehand ultrasound reconstruction. *Computer Methods and Programs in Biomedicine*, 211:106421, 2021.
- [33] Andrew H Gee, R James Housden, Peter Hassenpflug, Graham M Treece, and Richard W Prager. Sensorless freehand 3D ultrasound in real tissue: Speckle decorrelation without fully developed speckle. *Medical Image Analysis*, 10(2):137–149, 2006.
- [34] Loïc Tetrel, Hacène Chebrek, and Catherine Laporte. Learning for graph-based sensorless freehand 3D ultrasound. In *Machine Learning in Medical Imaging: 7th International Workshop, MLMI 2016, Held in Conjunction with MICCAI 2016, Athens, Greece, October 17, 2016, Proceedings 7*, pages 205–212. Springer, 2016.
- [35] Pierrick Coupé, Pierre Hellier, Xavier Morandi, and Christian Barillot. Probe trajectory interpolation for 3D reconstruction of freehand ultrasound. *Medical Image Analysis*, 11(6):604–615, 2007.
- [36] Tiexiang Wen, Feng Yang, Jia Gu, and Lei Wang. A novel bayesian-based nonlocal reconstruction method for freehand 3D ultrasound imaging. *Neurocomputing*, 168:104–118, 2015.
- [37] Raúl San José-Estépar, Marcos Martín-Fernández, P. Pablo Caballero-Martínez, Carlos Alberola-López, and Juan Ruiz-Alzola. A theoretical framework to three-dimensional ultrasound reconstruction from irregularly sampled data. *Ultrasound in Medicine & Biology*, 29(2):255–269, 2003.
- [38] Christoph B Burckhardt. Speckle in ultrasound b-mode scans. *IEEE Transactions on Sonics and Ultrasonics*, 25(1):1–6, 1978.

- [39] DEO Dewi, MHF Wilkinson, TLR Mengko, IKE Purnama, PMA Van Ooijen, AG Veldhuizen, NM Maurits, and GJ Verkerke. 3D ultrasound reconstruction of spinal images using an improved olympic hole-filling method. In *International Conference on Instrumentation, Communication, Information Technology, and Biomedical Engineering 2009*, pages 1–5. IEEE, 2009.
- [40] QH Huang and YP Zheng. A new scanning approach for limb extremities using a water bag in freehand 3D ultrasound. *Ultrasound in Medicine & Biology*, 31(4):575–583, 2005.
- [41] Qing-Hua Huang and Yong-Ping Zheng. Volume reconstruction of freehand three-dimensional ultrasound using median filters. *Ultrasonics*, 48(3):182–192, 2008.
- [42] Qinghua Huang, Yanping Huang, Wei Hu, and Xuelong Li. Bezier interpolation for 3D freehand ultrasound. *IEEE Transactions on Human-Machine Systems*, 45(3):385–392, 2014.
- [43] Wilna Du Toit. *Radial basis function interpolation*. PhD thesis, Stellenbosch: Stellenbosch University, 2008.
- [44] Robert Rohling, Andrew Gee, Laurence Berman, and Graham Treece. Radial basis function interpolation for freehand 3D ultrasound. In *Information Processing in Medical Imaging: 16th International Conference*, pages 478–483. Springer, 1999.
- [45] Raphael Prevost, Mehrdad Salehi, Julian Sprung, Alexander Ladikos, Robert Bauer, and Wolfgang Wein. Deep learning for sensorless 3D freehand ultrasound imaging. In *International Conference on Medical Image Computing and Computer-assisted Intervention*, pages 628–636. Springer, 2017.
- [46] Mingliang Zhai, Xuezhi Xiang, Ning Lv, and Xiangdong Kong. Optical flow and scene flow estimation: A survey. *Pattern Recognition*, 114:107861, 2021.
- [47] Narges Afsham, Mohammad Najafi, Purang Abolmaesumi, and Robert Rohling. A generalized correlation-based model for out-of-plane motion estimation in freehand ultrasound. *IEEE*, 33(1):186–199, 2013.
- [48] R James Housden, Andrew H Gee, Graham M Treece, and Richard W Prager. Sensorless reconstruction of unconstrained freehand 3D ultrasound data. *Ultrasound in Medicine & biology*, 33(3):408–419, 2007.
- [49] Catherine Laporte and Tal Arbel. Combinatorial and probabilistic fusion of noisy correlation measurements for untracked freehand 3D ultrasound. *IEEE Transactions on Medical Imaging*, 27(7):984–994, 2008.

- [50] Ruey-Feng Chang, Wen-Jie Wu, Dar-Ren Chen, Wei-Ming Chen, Wei Shu, Jau-Hong Lee, and Long-Bin Jeng. 3D US frame positioning using speckle decorrelation and image registration. *Ultrasound in Medicine & Biology*, 29(6):801–812, 2003.
- [51] Anirudha Ghosh, Abu Sufian, Farhana Sultana, Amlan Chakrabarti, and Debashis De. Fundamental concepts of convolutional neural network. pages 519–567, 2020.
- [52] Kanta Miura, Koichi Ito, Takafumi Aoki, Jun Ohmiya, and Satoshi Kondo. Localizing 2D ultrasound probe from ultrasound image sequences using deep learning for volume reconstruction. In *Medical Ultrasound, and Preterm, Perinatal and Paediatric Image Analysis: First International Workshop, ASMUS 2020, and 5th International Workshop, PIPPI 2020*, pages 97–105. Springer, 2020.
- [53] Moez Krichen. Convolutional neural networks: a survey. *Computers*, 12(8):151, 2023.
- [54] Wolfgang Wein, Mattia Lupetti, Oliver Zettinig, Simon Jagoda, Mehrdad Salehi, Viktoria Markova, Dornoosh Zonoobi, and Raphael Prevost. Three-dimensional thyroid assessment from untracked 2D ultrasound clips. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 514–523. Springer, 2020.
- [55] Hengtao Guo, Sheng Xu, Bradford Wood, and Pingkun Yan. Sensorless freehand 3D ultrasound reconstruction via deep contextual learning. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020*, pages 463–472. Springer, 2020.
- [56] Hengtao Guo, Sheng Xu, Bradford J Wood, and Pingkun Yan. Transducer adaptive ultrasound volume reconstruction. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 511–515. IEEE, 2021.
- [57] Hengtao Guo, Hanqing Chao, Sheng Xu, Bradford J Wood, Jing Wang, and Pingkun Yan. Ultrasound volume reconstruction from freehand scans without tracking. *IEEE Transactions on Biomedical Engineering*, 70(3):970–979, 2022.
- [58] Qi Li, Ziyi Shen, Qian Li, Dean C Barratt, Thomas Dowrick, Matthew J Clarkson, Tom Vercauteren, and Yipeng Hu. Trackerless freehand ultrasound with sequence modelling and auxiliary transformation over past and future frames. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2023.
- [59] Hojjat Salehinejad, Sharan Sankar, Joseph Barfett, Errol Colak, and Shahrokh Valaee. Recent advances in recurrent neural networks. *ArXiv*, 2017.

- [60] Alex Sherstinsky. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *Physica D: Nonlinear Phenomena*, 404:132306, 2020.
- [61] Yuxiu Hua, Zhifeng Zhao, Rongpeng Li, Xianfu Chen, Zhiming Liu, and Honggang Zhang. Deep learning with long short-term memory for time series prediction. *IEEE Communications Magazine*, 57(6):114–119, 2019.
- [62] Liang Gonog and Yimin Zhou. A review: generative adversarial networks. In *2019 14th IEEE Conference on Industrial Electronics and Applications (ICIEA)*, pages 505–510. IEEE, 2019.
- [63] Hamed Alqahtani, Manolya Kavakli-Thorne, and Gulshan Kumar. Applications of generative adversarial networks (GANs): an updated review. *Archives of Computational Methods in Engineering*, 28:525–552, 2021.
- [64] Jiannan Zheng, Shun Miao, Z Jane Wang, and Rui Liao. Pairwise domain adaptation module for cnn-based 2D/3D registration. *Journal of Medical Imaging*, 5(2):021204–021204, 2018.
- [65] Dwarikanath Mahapatra and Zongyuan Ge. Training data independent image registration using generative adversarial networks and domain adaptation. *Pattern Recognition*, 100:107109, 2020.
- [66] Mingyuan Luo, Xin Yang, Xiaoqiong Huang, Yuhao Huang, Yuxin Zou, Xindi Hu, Nishant Ravikumar, Alejandro F Frangi, and Dong Ni. Self context and shape prior for sensorless freehand 3D ultrasound reconstruction. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France*, pages 201–210. Springer, 2021.
- [67] Thomas Leblanc, Florent Lalys, Quentin Tollenaere, Adrien Kaladji, Antoine Lucas, and Antoine Simon. Stretched reconstruction based on 2D freehand ultrasound for peripheral artery imaging. *International Journal of Computer Assisted Radiology and Surgery*, 17(7):1281–1288, 2022.
- [68] Yimeng Dou, Fangzhou Mu, Yin Li, and Tomy Varghese. Sensorless end-to-end freehand three-dimensional ultrasound reconstruction with physics guided deep learning. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, 2024.
- [69] Jana Zujovic, Thrasyvoulos N Pappas, and David L Neuhoff. Structural texture similarity metrics for image analysis and retrieval. *IEEE Transactions on Image Processing*, 22(7):2545–2558, 2013.
- [70] Alain Hore and Djemel Ziou. Image quality metrics: PSNR vs. SSIM. In *2010 20th international conference on pattern recognition*, pages 2366–2369. IEEE, 2010.

- [71] Umme Sara, Morium Akter, and Mohammad Shorif Uddin. Image quality assessment through FSIM, SSIM, MSE and PSNR - a comparative study. *Journal of Computer and Communications*, 7(3):8–18, 2019.
- [72] Wufeng Xue, Lei Zhang, Xuanqin Mou, and Alan C Bovik. Gradient magnitude similarity deviation: A highly efficient perceptual image quality index. *IEEE Transactions on Image Processing*, 23(2):684–695, 2013.
- [73] Narasimhan Venkatanath, D Praneeth, Maruthi Chandrasekhar Bh, Sumohana S Channappayya, and Swarup S Medasani. Blind image quality evaluation using perception based features. In *2015 twenty-first national conference on communications (NCC)*, pages 1–6. IEEE, 2015.
- [74] Dominik Müller, Iñaki Soto-Rey, and Frank Kramer. Towards a guideline for evaluation metrics in medical image segmentation. *BMC Research Notes*, 15(1):210, 2022.
- [75] Amin Pourmahboubi, Nazanin Arsalani Saeed, and Hamed Tabrizchi. A brain tumor segmentation enhancement in mri images using u-net and transfer learning. *BMC Medical Imaging*, 25(1):307, 2025.
- [76] Varduhi Yeghiazaryan and Irina Voiculescu. Family of boundary overlap metrics for the evaluation of medical image segmentation. *Journal of Medical Imaging*, 5(1):015006–015006, 2018.
- [77] FUJIFILM VisualSonics. *Vevo 3100 user manual*, 2017.
- [78] Northern Digital Inc. Aurora field generators – electromagnetic tracking technology, n.d. Available at: <https://www.ndigital.com/electromagnetic-tracking-technology/aurora/>. [Accessed: 01-August-2025].
- [79] Plus Toolkit. Plus applications user manual- temporal calibration algorithm. Available at: <http://perk-software.cs.queensu.ca/plus/doc/nightly/user/AlgorithmTemporalCalibration.html>. [Accessed: 09-August-2025].
- [80] João Luís Gomes da Fonseca. *Surgical Navigation System for Percutaneous Renal Access: electromagnetic guidance and imaging working together*. Phd thesis in health sciences, School of Medicine, University of Minho, Braga, Portugal, 2021.
- [81] F Landis Markley, Yang Cheng, John L Crassidis, and Yaakov Oshman. Quaternion averaging. *Journal of Guidance, Control, and Dynamics*, 30, 2007.

- [82] João Gomes-Fonseca, Fernando Veloso, Sandro Queirós, Pedro Morais, António CM Pinho, Jaime C Fonseca, Jorge Correia-Pinto, Estêvão Lima, and João L Vilaça. Assessment of electromagnetic tracking systems in a surgical environment using ultrasonography and ureteroscopy instruments for percutaneous renal access. *Medical physics*, 47(1):19–26, 2020.
- [83] Gregorio Andria, Filippo Attivissimo, Attilio Di Nisio, Anna Maria Lucia Lanzolla, and Mattia Alessandro Ragolia. Analysis and optimization of surgical electromagnetic tracking systems by using magnetic field gradients. *Acta IMEKO*, 12(2):1–8, 2023.
- [84] Tapani Koivukangas, Jani PA Katisko, and John P Koivukangas. Technical accuracy of optical and the electromagnetic tracking systems. *SpringerPlus*, 2(1):90, 2013.
- [85] Qi Li, Shaheer U Saeed, Yuliang Huang, Mingyuan Luo, Zhongnuo Yan, Jiongquan Chen, Xin Yang, Dong Ni, Nektarios Winter, Phuc Nguyen, et al. Tus-rec2024: A challenge to reconstruct 3d freehand ultrasound without external tracker. *arXiv preprint arXiv:2506.21765*, 2025.
- [86] Van-Thanh Hoang and Kang-Hyun Jo. Practical analysis on architecture of efficientnet. In *2021 14th International Conference on Human System Interaction (HSI)*, pages 1–4. IEEE, 2021.
- [87] Jiang Shao, Hanbao Wu, and Hao Liu. Fine-grained recognition based on improved efficientnet model. In *2021 IEEE International Conference on Unmanned Systems (ICUS)*, pages 585–589. IEEE, 2021.
- [88] Xiangchen Yao, Shuqi Ma, Zhaoyuan Zhang, Yuanzhen Xu, Ziyi Bai, and Bo Li. Robust leakage detection in tunnels with unstable illumination using an explainable deep learning approach. *Transportation Infrastructure Geotechnology*, 12(3):1–28, 2025.
- [89] Ammar Mohammed and Rania Kora. A comprehensive review on ensemble deep learning: Opportunities and challenges. *Journal of King Saud University-Computer and Information Sciences*, 35(2):757–774, 2023.

Appendix A

Design of the sensor holder

This appendix presents the complete sensor holder setup, including both the SolidWorks design and the final 3D-printed version. It is important to note that the entire probe–holder system was modeled in SolidWorks, as no official 3D model of the probe was available. Consequently, to design a holder that would precisely fit the probe, the probe itself first had to be recreated in the software, as shown in Figure 42. This process required multiple physical measurements using a caliper, enabling the probe geometry to be modeled as accurately as possible. The holder was then designed around this reconstructed probe geometry to ensure a stable and secure fit (Figure 43). The complete probe–holder system with the sensor attached is shown in Figure 44. Finally, a comparison between the SolidWorks design of the holder and its final 3D-printed implementation attached to the MX250 probe is presented in Figure 45.

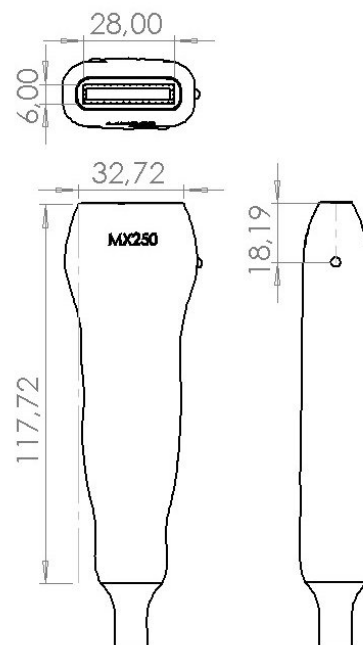


Figure 42: Technical drawing of the MX250 probe.

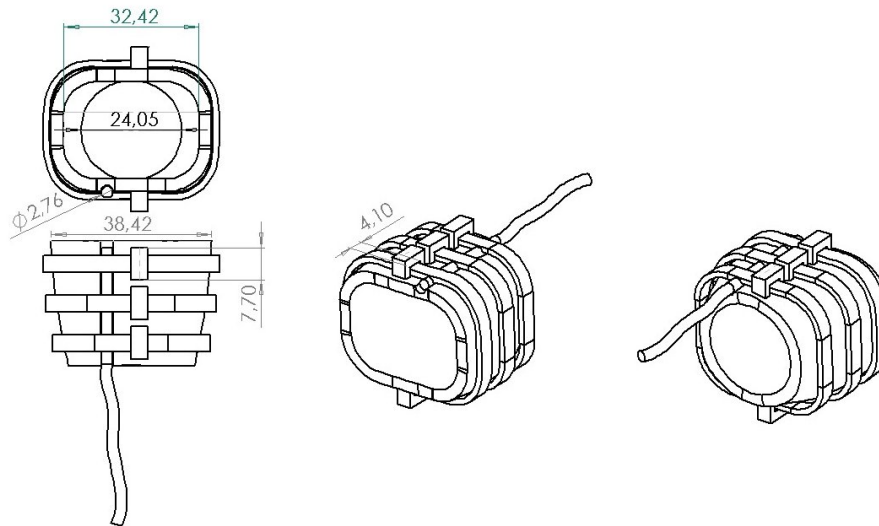


Figure 43: Technical drawing of the holder designed to secure the sensor to the probe.

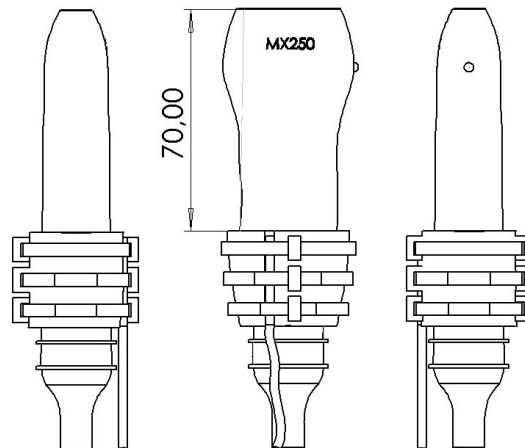


Figure 44: Technical drawing of the MX250 probe showing the sensor attached to the holder.



Figure 45: Sensor holder design in SolidWorks (left) and its real 3D-printed implementation attached to the MX250 probe (right).

Appendix B

CAD schematics for tracker–US validation

This appendix presents the technical drawings created in SolidWorks, illustrating all relevant dimensions considered in the validation setup. It is important to note that the validation setup was developed through a trial-and-error process, as the exact probe model was not available.

As previously mentioned, ensuring that the probe consistently fitted into the same position and remained fully immobilized during each validation acquisition was essential. The overall geometry of the support structure, including the probe holder and fixation elements, is detailed in Figure 46. As shown in this figure, a dedicated cavity was designed to accommodate the probe, together with a small ramp at the front of the support structure to guide the probe into the correct position and ensure stable and repeatable placement across acquisitions.

The complete validation setup, including the support structure placed inside the water container, is illustrated in Figure 47. Additional CAD views of the validation model are provided in Figure 48, showing the front, isometric, and side perspectives of the structure. The assembled setup inside the water container is further illustrated in Figure 49, where different views show the acrylic cover used to keep the model partially submerged during the acquisitions.

Finally, the internal wire configuration used for the accuracy assessment is shown in Figure 50. These drawings present the wires tensioned and arranged along straight, parallel paths between opposing points of the structure, which served as fixed reference lines during the validation experiments.

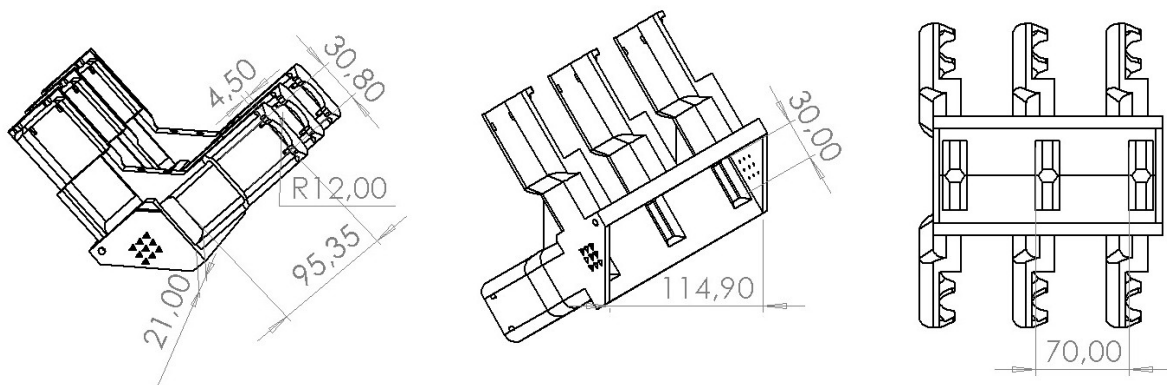


Figure 46: Technical drawings of the validation support structure.

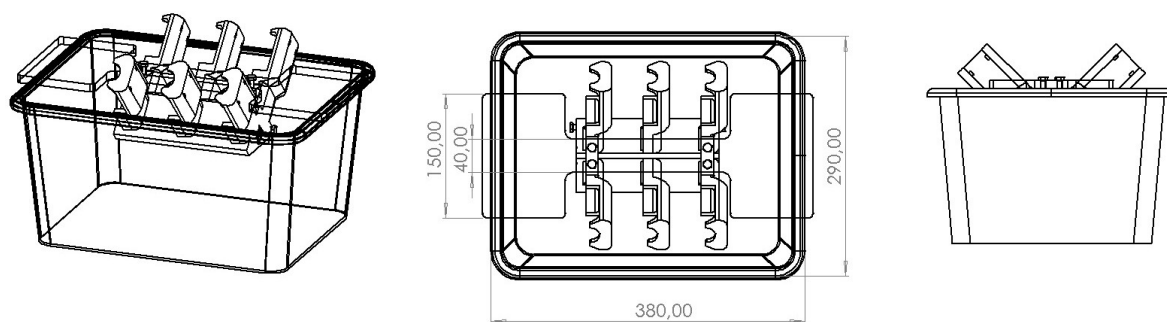


Figure 47: Technical drawings of the custom validation setup developed for this study.

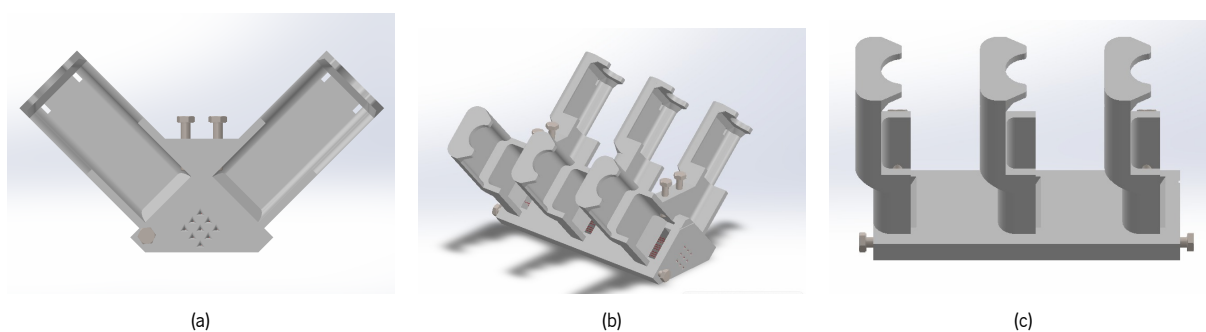


Figure 48: CAD views of the validation model showing (a) front, (b) isometric, and (c) side perspectives.

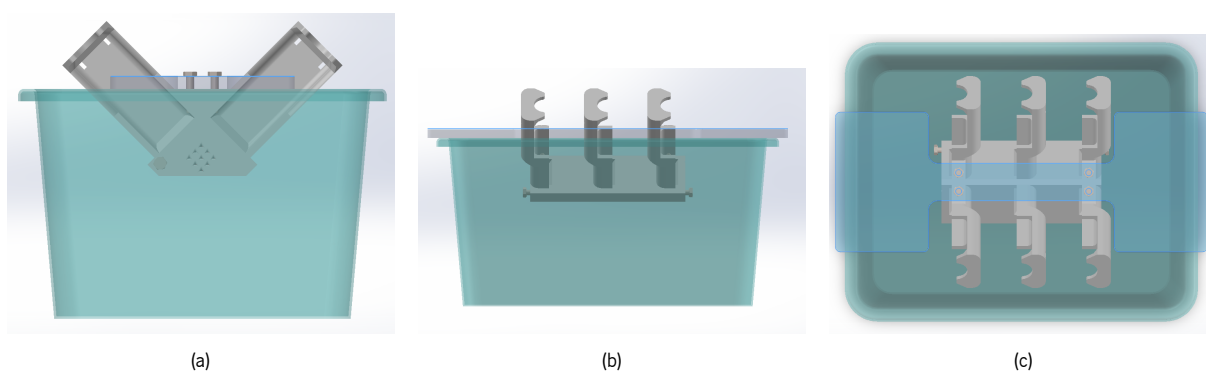


Figure 49: Different views of the validation model inside the water container, with an acrylic cover used to keep the model partially submerged: (a) front view, (b) side view, and (c) top view.

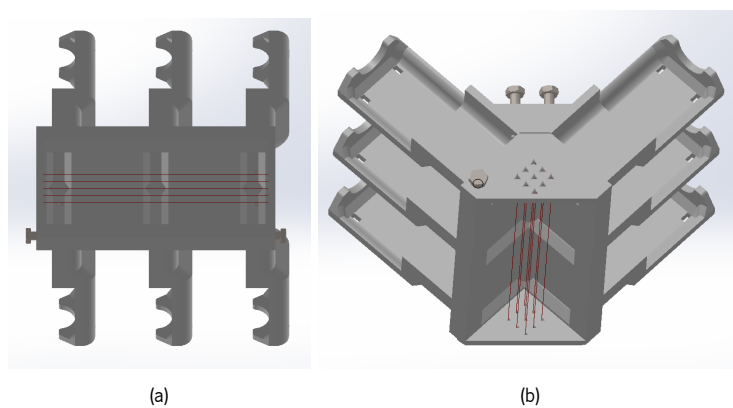


Figure 50: Internal wire configuration of the validation model: (a) top view and (b) isometric view.