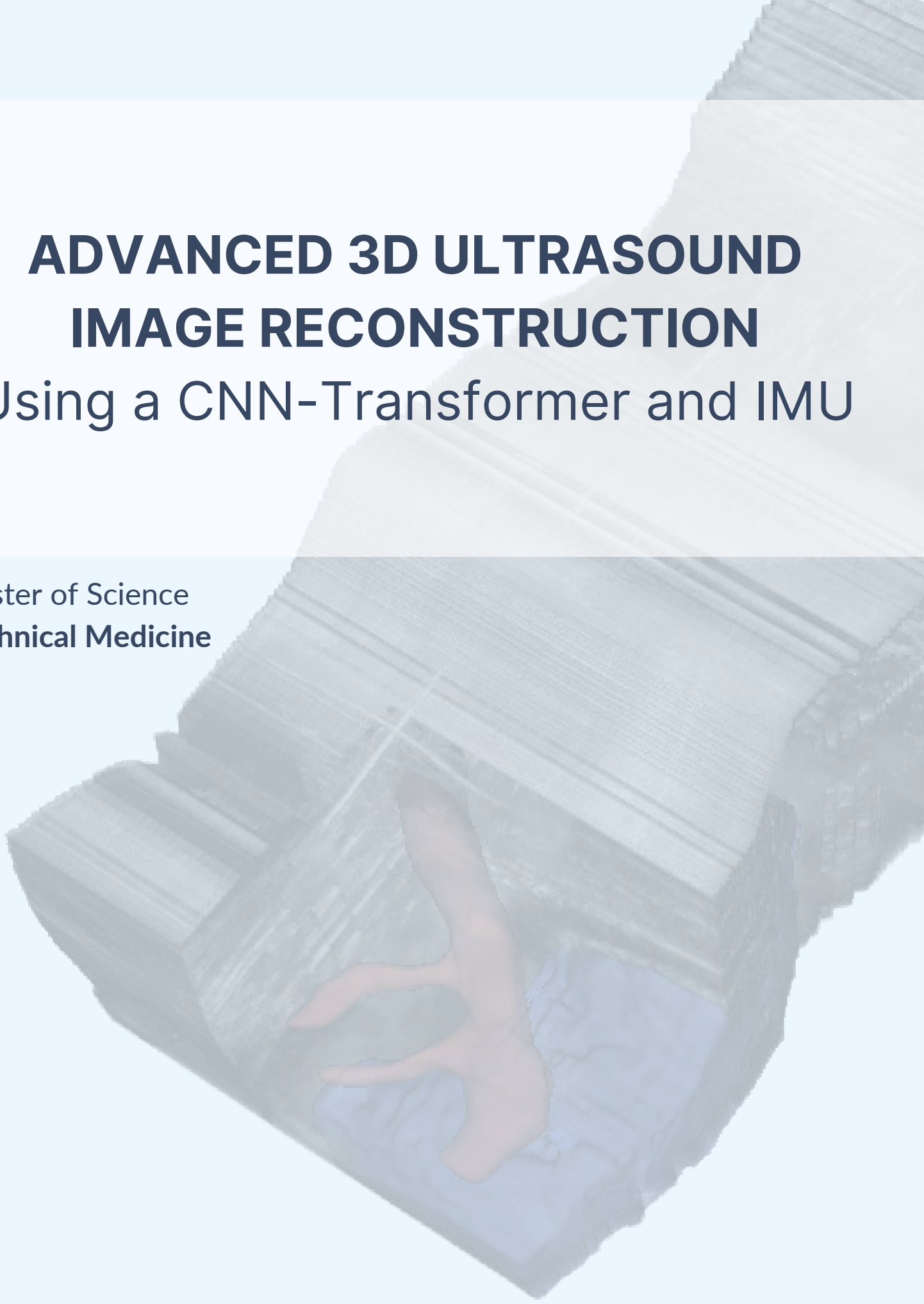
ADVANCED 3D ULTRASOUND IMAGE RECONSTRUCTION

Using a CNN-Transformer and IMU



Master of Science
Technical Medicine

Chrissy A. Adriaans
November 11, 2024

Advanced 3D Ultrasound Image Reconstruction Using a CNN-Transformer and IMU

Chrissy A. Adriaans
Student number : 4560957
November 11, 2024

Thesis in partial fulfilment of the requirements for the joint degree of Master of Science in

Technical Medicine

Leiden University ; Delft University of Technology ; Erasmus University Rotterdam

Master thesis project (TM30004 ; 35 ECTS)

Dept. of Biomechanical Engineering, TU Delft
19-02-2024 – 23-09-2024

Supervisor(s):

Dr. Freija Geldof
Dr. Behdad Dashtbozorg
Prof. Dr. Jos. A. van der Hage

Thesis committee members:

Dr. Jifke F. Veenland, Erasmus MC (chair)
Dr. Freija Geldof, NKI-AvL
Dr. Behdad Dashtbozorg, NKI-AvL
Prof. Dr. Jos. A. van der Hage, LUMC

Universiteit Leiden

TU Delft

Erasmus Universiteit Rotterdam

Preface

This thesis marks the culmination of my Master's studies in Technical Medicine, as well as the research I conducted within the Surgical Oncology department at the AvL-NKI. Working on this project has been challenging but mostly rewarding, as I had the chance to leverage my knowledge gained throughout my studies and explore new areas, from building a robotic setup to diving into sensor systems.

I would like to express my special gratitude to my supervisors, Freija and Behdad, for their invaluable guidance throughout this project. Their support, dedication and availability for brainstorming played an important role in my enthusiasm for this project. Thank you for helping me tackle complex challenges, some of which we can now look back on and laugh. I'm also grateful to Jos for his medical insights and enthusiasm; hopefully this technology will find its way into clinical practice. Of course, a huge thanks for the weekly POCUS meetings as well, with Mark, Lennard, Hilde, and Lucie. I really enjoyed my time at the NKI, and might even miss the POCUS meetings in a few weeks. Already looking forward to the little reunion in December!

This thesis also marks the end of my time as a student, a period filled with joy that began in Delft, continued in Rotterdam, and finally Amsterdam. To my friends, family, and housemates, I thank you, not only for your support but also for all the fun distractions along the way!

As I take the next step forward, I would like to thank you for taking the time to read my thesis. I hope it provides valuable insights and perhaps even sparks new ideas for future exploration.

Chrissy Adriaans
Amsterdam, November 2024

Abstract

Introduction. Three-dimensional (3D) ultrasound (US) offers significant potential to enhance diagnostic imaging or intraoperative guidance by providing comprehensive volumetric insights in a non-invasive and cost-effective manner. However, existing methods for 3D reconstruction often rely on external tracking devices or specialized 3D transducers, which are costly and less suited for intraoperative and point-of-care settings.

Aim. This thesis aims to advance trackerless 3D US reconstruction, leveraging a point-of-care handheld ultrasound (POCUS) probe integrated with an inertial measurement unit (IMU) and a novel deep learning architecture.

Method. A high-quality dataset was acquired using a custom setup that included a POCUS probe with an integrated IMU, facilitating precise positional tracking and controlled movement of complex random motion trajectories. A CNN-Transformer network was developed, leveraging 2D US images and optical flow, to predict both local and global transformation parameters utilized for 3D US reconstruction. Ablation experiments were conducted to optimize model performance.

Results. A dataset comprising 361 US sweeps from ex-vivo surgical specimens and phantom models was collected. The optimized model, integrating IMU orientation data and applying sequence augmentation, achieved a mean Final Drift Ratio (FDR) of 11.63 ± 8.63% on an unseen test set, with a median FDR of 8.11%. Quantitative evaluations demonstrated that the model accurately captured the shape of the sweeps, particularly in translations along the x-axis, y-axis, and rotation around the x-axis. The predicted reconstructions enabled correct segmentation and visualization of anatomical structures in 3D, crucial for application in clinical settings.

Conclusion. This thesis presents a novel approach for 3D US reconstruction without external tracking devices, utilizing an integrated IMU and a CNN-Transformer network. The results demonstrate competitive performance with state-of-the-art methods, highlighting the feasibility and potential of this approach for applications in diagnostic imaging, surgical planning, and intraoperative guidance, advancing 3D US toward clinical integration and improved patient outcomes.

Contents

1 Introduction

Two-dimensional ultrasound (2D US) has established itself as an invaluable imaging modality in clinical settings due to its low cost, portability, and non-invasive nature. Its real-time visualization capability offers significant utility in intraoperative procedures, providing clinicians with essential anatomical and functional information. Despite these advantages, 2D US is hindered by limitations such as limited spatial orientation, lacking valuable 3D context, and its dependence on the operator's experience and interpretive skills. These constraints underscore the inherent variability and the potential for subjectivity in 2D US [1]. The introduction of three-dimensional ultrasound (3D US) marks a significant advancement, addressing many of these limitations. Specifically, 3D US provides a more comprehensive visualization of anatomical regions of interest (ROIs) and provides the flexibility of examining the acquired data from multiple viewpoints post-acquisition, reducing the variability due to operators. Moreover, this technology could democratize access to advanced imaging techniques, by seamlessly integrating with other 3D imaging modalities such as CT or MRI, providing a comprehensive view of patient anatomy and enabling more personalized and effective treatment plans.

With these advancements, 3D US extends the potential of 2D US across a broader spectrum of clinical applications and has the potential to revolutionize fields such as surgical planning and diagnostic imaging [2, 3]. Providing real-time, detailed 3D visualizations could enhance the precision of surgical procedures, reduce operation times, and improve patient outcomes, which can be illustrated by the assessment of tumor margins in oncological surgery. While histology results offer delayed feedback regarding resection margins, 3D US can provide real-time assessment during the surgical procedure. A clinical trial has demonstrated that intraoperative 3D US is an accurate method for assessing ex vivo surgical margins during tongue cancer surgery [4]. Consequently, the implementation of this technology may ultimately lead to an improvement in the number of free margins achieved after cancer treatment. In diagnostic imaging, reconstructing 3D structures facilitates quantitative volume measurements. This is particularly relevant in assessing prostate volume, a critical factor for diagnosing and managing prostate cancer, where current methods often lack sufficient accuracy [5]. Enhanced volume measurements, utilizing 3D US, may contribute to earlier detection and more effective treatment monitoring. This could be achieved in a non-invasive, low-cost, point-of-care setting, making advanced diagnostics more accessible, particularly in primary-care environments [6].

The path to realizing 3D US reconstructions has seen various methodologies, including the use of 3D US transducers, external tracking, and image-based reconstruction methods. 3D US transducers employ 2D crystal arrays rather than traditional 1D crystal arrays. This design facilitates the electronic steering and focusing of the US beam in multiple dimensions, thereby enabling the real-time

acquisition of 3D volumetric data [7]. However, since the transducer elements are spread across two dimensions, this configuration results in fewer elements in each dimension. Therefore, to maintain a sufficient spatial resolution the imaging system must limit the coverage area, resulting in a smaller field of view (FOV), particularly at greater depths. External tracking devices such as mechanical, optical, or electromagnetic (EM) systems are designed to position the US probe in 3D space. These systems are prone to artifacts caused by magnetic interference or optical occlusion and have constraints in the range of motion. Considering the high costs and cumbersome set-up of 3D US transducers and external tracking systems, their suitability for certain clinical applications like intraoperative and point-of-care US (POCUS) is diminished [8, 9].

Image-based freehand 3D US represents a promising alternative, eliminating the need for both external tracking devices and 3D transducers, thereby capitalizing on the potential for integration with cost-effective, portable and sometimes wireless, US devices [10]. The computational challenge of these systems lies in the reconstruction process, which necessitates an accurate estimation of the probe's trajectory, divisible into in-plane and out-of-plane (elevational) movements. While in-plane movements are more readily quantified, out-of-plane estimations remain complex. Prior research on out-of-plane motion dates back to seminal work by Chen *et al.* [11] and has been mainly based on speckle noise, the granular gray-scale textures in B-mode US images. Speckle decorrelation methods map the transformation between neighboring US images to the correlation of their speckle patterns, i.e. the higher the speckle correlation, the lower the elevational distance between neighboring frames [12]. Under Rayleigh scattering conditions, where the size of these scatters is much smaller than the wavelength of the sound waves, this speckle pattern is theoretically predictable. However, when applied to dynamic clinical scenarios, these models often fall short, as speckle variability and real tissue movement introduce errors that accumulate, leading to drift and a compromise in accuracy [13–15].

Recent advancements in artificial intelligence (AI), especially deep learning (DL) techniques, have expanded the horizons for extracting detailed information from image data [16]. This has enabled significant progress in overcoming the intrinsic difficulties of image-based 3D US reconstruction. DL networks can automatically learn to identify and prioritize essential features within the US images by leveraging labeled transformations associated with each image. This data-driven approach enables the network to capture complex spatial and temporal patterns inherent in the imaging data, which is crucial for accurately estimating probe trajectories and enhancing the precision and efficiency of the reconstruction process. While traditional tracking technologies that require external reference are less suited to point-of-care and intraoperative settings, inertial measurement units (IMUs) have emerged as a viable and less obtrusive alternative for position tracking [17]. These devices integrate a tri-directional magnetometer, a gyroscope, and an accelerometer into compact units. This offers a favorable balance between hardware independence and the need for positional information in trajectory reconstruction. With the growing adoption of IMU technology in clinical devices and developments in DL techniques, image-based reconstruction methods without external tracking are becoming increasingly viable. Furthermore, leveraging this IMU technology aligns with the ongoing shift towards more accessible and efficient POCUS applications.

Given the rapid developments in the novel field of trackerless 3D US reconstruction, a literature review was conducted to systematically evaluate the current state-of-the-art methodologies [18]. This review revealed that studies employing convolutional neural networks (CNNs), sometimes combined with IMU data, have shown promising results in refining freehand 3D reconstruction techniques compared to traditional approaches. However, several challenges persist, associated with dataset characteristics, reconstruction accuracy, and clinical applicability, highlighting the necessity for continued research in this field. Although this literature review showed the importance of variance in trajectories and anatomy, only 4 out of 23 datasets reported on sweeps encompassing different types of motion like angular movements or wave-shaped trajectories, affecting model training and generalization capabilities. Moreover, half of the datasets were confined to phantoms or arms, which may not adequately

represent the variability and challenges encountered in clinical practice.

These findings highlight the need for more comprehensive datasets and advanced modeling approaches to improve reconstruction accuracy and clinical applicability. Therefore, the aim of this thesis is twofold. First, it aims to acquire a high-quality dataset suitable for training a 3D US reconstruction model, utilizing a POCUS probe with an integrated IMU. Although this methodology can be adapted to a variety of clinical applications, this thesis specifically emphasizes data acquisition from resected ex vivo tumor specimens. This dataset aims to encompass a variety of tissue types and a wide range of sweep trajectories, including not only linear but also complex motions. The built-in IMU of the POCUS probe is leveraged to enhance reconstruction accuracy, eliminating the need for manual sensor mounting and enhancing clinical feasibility.

Secondly, this thesis develops a novel deep learning model with a CNN-Transformer architecture. While previous studies have employed CNNs with various strategies to capture temporal dependencies, Transformers represent the state-of-the-art in sequence modeling and have yet to be fully explored in this context [19]. By integrating a Transformer, the model aims to capture long-range dependencies within the data more effectively, possibly enhancing reconstruction accuracy.

In conclusion, this work seeks to lay the foundation for a workflow that can be leveraged for several clinical applications. By advancing trackerless 3D US reconstruction, this research aims to bring this technology one step closer to integration into routine clinical practice.

2

Methods

This chapter outlines the methodologies employed in this thesis, which are divided into three primary sections. Section 2.1 details the experimental design, including how a dataset of sweeps of 2D ultrasound images was collected, along with the corresponding inertial and positional tracking data. Consequently, all required data preparatation and preprocessing steps are discussed in Section 2.2. Finally, Section 2.3 details the development and training of a deep learning model for 3D trajectory reconstruction utilizing the collected dataset, encompassing the different methodologies employed for model optimization and evaluation.

2.1. Experimental design

The following section outlines the experimental design, encompassing the patient inclusion criteria and the hardware used in the data acquisition setup. Subsequently, the data collection process is described, comprising of the employed scanning procedures, specifications of the trajectories and the method by which the US images, inertial data and positional data are extracted to form the dataset.

2.1.1. Patient inclusion criteria

To develop a comprehensive dataset feasible for 3D reconstruction of 2D US images, data were acquired from two primary sources: a variety of ex-vivo specimens and a US phantom. This approach ensured variability in anatomical structures and tissue properties, enhancing the robustness and generalizability of the deep learning model.

Ex-vivo specimens were obtained directly after surgical resection procedures at the Netherlands Cancer Institute (NKI) in Amsterdam, including mastectomies, lumpectomies, colorectal surgeries, and sarcoma excisions. Both benign and malignant tissues were included to capture a spectrum of pathological and normal tissue characteristics. Immediate post-resection imaging of these specimens preserved tissue integrity and acoustic properties, providing realistic data for model training. Specimens that were too large to fit within the scanning setup (greater than 25 cm) or too small to provide sufficient imaging data (less than 4 cm) were excluded to ensure equipment compatibility. Specimens with excessive fluids or leakage that could pose a risk to equipment safety or tissue integrity were also not included. All patients gave their informed consent to participate in scientific research, ensuring compliance with ethical standards and patient confidentiality. Patient anonymity and data confidentiality were strictly maintained throughout the study.

In addition to ex vivo specimens, an abdominal intraoperative and laparoscopic ultrasound Phan-

tom 'IOUSFAN' was used (Kyoto Kagaku Co., Ltd., Kyoto, Japan). This phantom simulates detailed abdominal anatomy and is designed for training of intraoperative and laparoscopic US procedures. It offers a controlled environment to collect data that mimics the acoustic properties of human tissue.

2.1.2. Hardware of the acquisition setup

A custom data acquisition setup was designed to facilitate precise tracking of the US probe during scanning. The configuration of this setup was critical for obtaining accurate ground truth positional data necessary for training the 3D reconstruction model. By integrating the US probe with a motorized scanner, controlled movements and synchronization between imaging and positional data were achieved.

Point-Of-Care Ultrasound (POCUS) probe

The Clarius HD3 L20 was used for the acquisition of 2D US images. This is a wireless handheld POCUS scanner equipped with an internal 9-degree-of-freedom (9-DOF) IMU sensor. The Clarius HD3 L20 features a 25 mm wide field of view and operates within a frequency range of 8–20 MHz, with a maximum imaging depth of 4 cm [20]. This device facilitates real-time streaming of both US images and IMU data, enabling synchronized acquisition potentially of great value for 3D reconstruction tasks. In this study, the probe operated in Research Mode, permitting customization of imaging parameters and enabling of the IMU sensor.

Ultrasound images

Images were captured in B-mode at an imaging depth of 2 cm, utilizing a frequency of 14 MHz and a frame rate ranging between 10 and 25 frames per second (fps), depending on the probe's temperature and battery life. The 2 cm imaging depth was chosen to visualize superficial structures such as tumor margins using higher frequencies, providing high-resolution images while maintaining a sufficient frame rate for accurate spatial reconstruction of the images during continuous probe movement.

Inertial Measurement Unit (IMU)

The integrated IMU comprises an accelerometer, a gyroscope, and magnetometer, providing comprehensive motion tracking capabilities that can be valuable for reconstructing 3D volumes from sequential 2D images. Each US frame was associated with multiple IMU data points, varying between one and three. The coordinate system of the IMU follows an East-North-Up configuration for the x, y and z axis, respectively. The accelerometer and gyroscope are positioned at an offset from the center of the imaging array of $(-107.50$ mm, -7.20 mm, 6.54 mm) along the x, y, and z axis. The magnetometer is located at $(-92.19$ mm, -21.73 mm, 6.46 mm). These spatial offsets are critical for accurate alignment between the IMU data, US images and ground truth position tracking during data processing. The orientation and offsets relative to the probe are illustrated in Figure 2.1(a).

Motorized scanner

To facilitate controlled movement of the US probe and obtain accurate ground truth positional data, a motorized scanner was employed to execute the US sweeps. The scanner was based on a modified NEJE laser engraver machine, from which the laser component was removed to accommodate the Clarius US probe. The original apparatus operated along two perpendicular linear axes (x and z), which were extended to increase the range of motion, allowing comprehensive coverage of the specimens. Linear motion along these axes was driven by two NEJE stepper motors, providing precise positioning relative to the starting point of each sweep on the horizontal plane (XZ). A manually adjustable y-axis was introduced to position the probe at the correct height above the specimen, ensuring optimal visualization.

To replicate the complex movements performed during manual scanning and to create a representative dataset, an additional degree of freedom (DOF) was added by incorporating a rotation about the lateral x-axis. This represented a tilting motion in the scanning direction, considered the most critical angular movement for this application. Introducing more DOF increases mechanical complexity and requires more sophisticated control algorithms. Therefore, a trade-off was made to balance representativeness and system complexity.

The original machine was modified to include this rotational DOF by integrating a third stepper motor connected to the NEJE motherboard via an additional rotational interface. A custom 3D-printed gearbox connected the framework of the motorized scanner, the US probe holder, and the third stepper motor, facilitating controlled tilting of the probe. The 3D-printed probe holder was designed to precisely fit the Clarius probe, eliminating any unwanted movement during scanning and ensuring consistent positioning across all sweeps. The rotation point was aligned with the midline of the probe and located at an offset of 82.5 mm from the imaging array. The final design enables automated movement along the x- and z-axes and rotational tilting, introducing small additional variations in the manual set height (y) of the image acquisition. This configuration provides realistic movements while maintaining the integrity of positional data, closely replicating manual scanning procedures. A schematic overview of the motorized scanner and its components is shown in Figure 2.1(b).

Detailed specifications of the NEJE machine, including the motherboard settings and stepper motor configurations, are provided in Appendix A. These settings were optimized to ensure smooth and fluent movements with minimal jerks or vibrations, which could adversely affect image quality or positional accuracy. Specifically, the step sizes were calibrated such that one step corresponds to 1 mm of linear motion along the x and z axes. The rotational step size for the tilting motion was configured to correspond to an angular increment of 0.038 degrees, determined through calibration experiments.

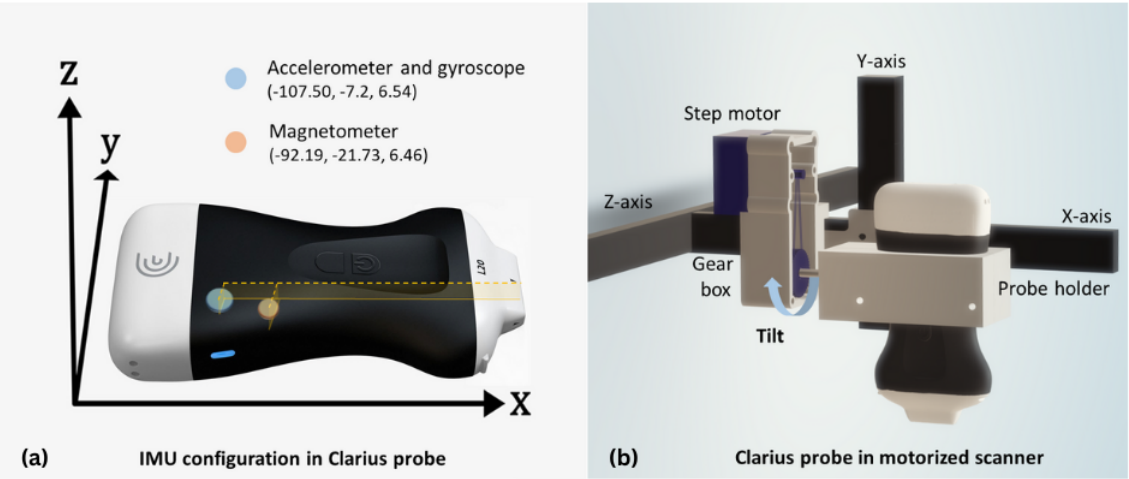


Figure 2.1: Schematic representation of (a) the Clarius HD3 L20 US probe illustrating the orientation of the IMU and locations of the accelerometer, gyroscope and magnetometer relative to the imaging array; (b) the US probe mounted in customized holder attached to gearbox and motorized scanner.

2.1.3. Data collection

Scanning procedure

To perform the US sweeps, the previously described hardware was utilized in a controlled scanning procedure designed to collect high-quality data for 3D reconstruction. Specimens were placed inside vacuum-sealed bags and submerged in a water-filled container to ensure optimal acoustic coupling, as the US probe was not in direct contact with the tissue to prevent movement and deformation. The use of vacuum bags allowed for secure mounting of the specimens using magnets, preventing movement due to floating during the scanning process. For the phantom measurements, the phantom was similarly immersed in water to facilitate acoustic contact. An illustration of this data acquisition setup is shown in Figure 2.2(a).

Multiple sweeps were performed per specimen to capture sufficient data across different areas. In order to maintain integrity and ensure variety of the dataset, the scanning trajectories did not show any overlap. Depending on the specimen size, the motorized scanner executed a number of predefined, randomly generated grids, following the criteria described in the next section. The scanner was programmed to move the probe along these grid coordinates at a consistent speed randomly selected within a range of 1.5 to 3 mm/s. Speed settings remained constant for all grids per specimen due to time efficiency, as the settings were configured on the scanner's motherboard. After each sweep, the probe was manually repositioned to a new start position corresponding to the end position of the executed sweep.

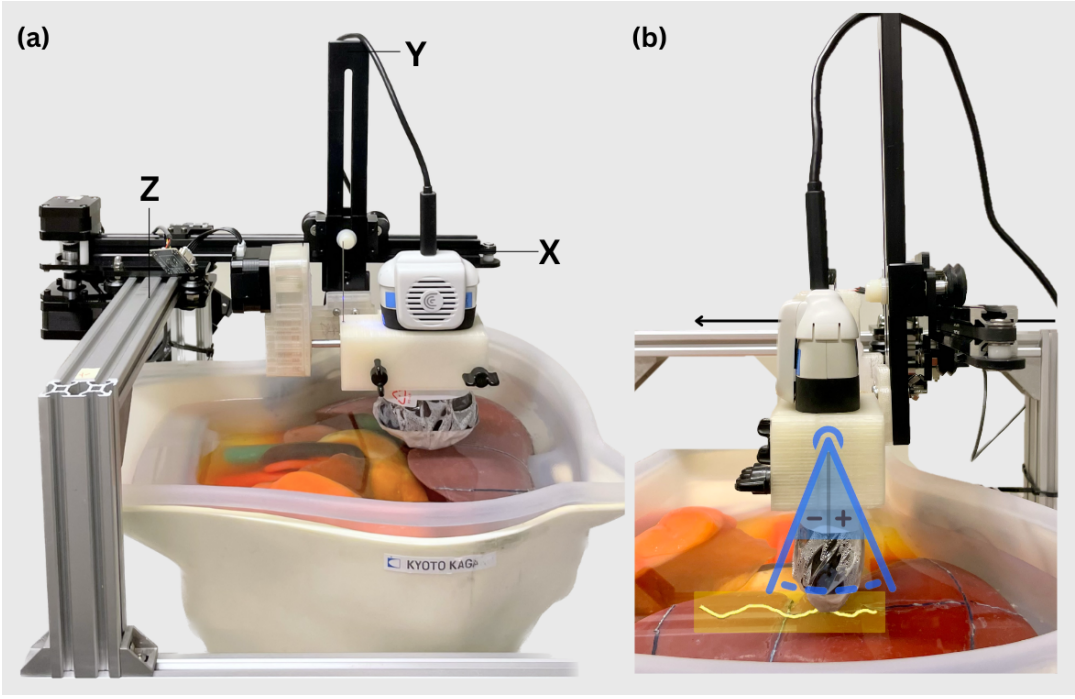


Figure 2.2: (a) Data acquisition setup with motorized scanner over water filled abdominal US phantom. (b) Side view with an example of a grid trajectory on the horizontal plane (XZ) in yellow and tilting range in blue.

Grid trajectory specifications

The grids were designed to simulate representative probe movements similar to those performed during manual scanning, introducing variability in scanning trajectories. Randomized trajectories constrained by predefined limits (e.g., maximum deviation on the x-axis and maximum length in the scanning direction (z-axis)) were created to mimic realistic motions while remaining within the physical capabilities

of the setup. The grids consisted of G-code commands driving the stepper motors, representing linear coordinate points (x, z) with associated tilt angles (θ_x) relative to the start position $(0, 0, 0)$.

The x-axis represented lateral movements, in which random step sizes between 0.05 mm and 1 mm were made, with direction changes introduced randomly to mimic natural scanning patterns. The z-axis corresponded to the scanning direction along the length of the specimen, in which random step sizes between 1 mm and 3 mm were made, ensuring consistent forward progression over the specimen. This approach generated a meandering path over the specimen, resembling the exploratory movements of manual scanning.

Tilt angles (θ_x) were incorporated to vary the orientation of the probe during scanning. This was established by generating sinusoidal functions with random frequencies, amplitudes, and phase shifts within set limits, resulting in a range of tilt angles from negative to positive values relative to the start position. The tilt angles were generated to start with negative values and progress to positive values, simulating the tilting motion a clinician might perform. In addition, random offsets were introduced to the sinusoidal trajectory to provide more variation in the tilt of the probe. An example of a generated scanning grid over the phantom is illustrated in Figure 2.2(b), in which the tilt motion is visualized by the range of angular variation facilitated by the grids.

Due to the offset between the probe's rotation point and its imaging array (82.5 mm), tilting the probe affects the imaging plane's position in both z (forward) and y (upward) directions. To account for this displacement and maintain a consistent forward motion of the imaging array, a correction factor was applied to the z-coordinate in the grid based on the tilt angle, using the trigonometric relationship:

$$\Delta z = 82.5 \cdot \cos\left(\frac{3\pi}{2} + \theta_x\right),$$

where Δz is the correction applied to the z-coordinate on the trajectory grid and θ_x is the angle of tilt in radians. This correction ensured that the imaging plane remained aligned with the intended scanning path, preventing excessive forward or backward shifts due to probe rotation.

Data extraction and synchronisation

The scanning process was controlled through a custom software developed in Python (version 3.8.18), that simultaneously executed the movement of the scanner, captured ground truth positional data, and streamed data from the Clarius probe. US images and IMU data of the Clarius were retrieved through Python using the Clarius Cast API (version 11.2.0), which requires the Clarius App (version 11.2.2) to be running simultaneously. Each frame and IMU data was tagged with a time of acquisition relative to the device startup time, designated as '*timestamp*'. The movement of the scanner was controlled using G-code programming, allowing precise control of the movement and facilitating the implementation of customized scanning trajectories. Simultaneously, the ground truth position was continuously extracted using serial communication, where the motors' responses to the G-code commands provided machine position data for each axis, along with machine status data. A detailed overview of the collected data is shown in Table 2.1.

To ensure synchronization of the data streams, the data acquisition of the Clarius probe was initiated first. A two-second delay was incorporated to ensure stable connections and synchronization, before movement of the scanner along the predefined grid was initiated. Upon completion of the movement, a 20-second delay was incorporated to account for data buffering, ensuring all data were received before saving. Since the data retrieved from Clarius provided timestamps relative to the device startup time, absolute time synchronization was achieved by aligning the initial timestamp from Clarius with the initial ground truth position, adjusted by the two-seconds advance of initialization. Wireless connections and buffering could introduce delays in data acquisition and retrieval from the Clarius device. By utilizing the USB serial connection for the ground truth data and carefully accounting for timing off-

Table 2.1: Dataset specifications

Acquisition Tool	Data Type	Measurements	Units	Temporal info
Clarius (Cast API)	US image	PNG image (480 × 640 pixels)	Grayscale	Timestamp
Clarius (Cast API)	Inertial data	- Accelerometer (x, y, z) - Gyroscope (x, y, z) - Magnetometer (x, y, z) - Orientation quaternion	- m/s^2 - rad/s - Gauss - Dimensionless	Timestamp
Motorized scanner (USB Serial)	Ground truth position	X, Y, Tilt positions relative to start	Stepper motor coordinates	Execution time (wall-clock)

sets, accurate synchronization between the datasets was achieved. This synchronization was critical for aligning the US images and inertial data with the corresponding positional ground truth, providing a comprehensive dataset for subsequent 3D reconstruction.

2.2. Data preparation and preprocessing

Data preprocessing is a crucial step in preparing the acquired dataset for developing a deep learning model for 3D trajectory reconstruction. The raw data, comprising US images, inertial data, and ground truth positional data, require alignment, synchronization, and formatting to ensure consistency and usability. This section first details the preparation required for the synchronized sweep trajectory data and the determination of the ground-truth US position labels for training of the model. Secondly, the preprocessing of the input data is outlined, including the US images, the computation of optical flow, and lastly the inertial data from the IMU measurements. This process ultimately provides the data that is input into the model.

2.2.1. Preparation of sweep trajectory

The preparation of the complete sweep of streamed data acquired during data collection, comprising US images and IMU data, was conducted in three stages. The initial stage involved alignment, followed by cropping and then subsampling to obtain the final datasets. Subsequently, ground truth position information was assigned to each frame in the dataset based on the movement of the motorized scanner. These steps are outlined in Figure 2.3.

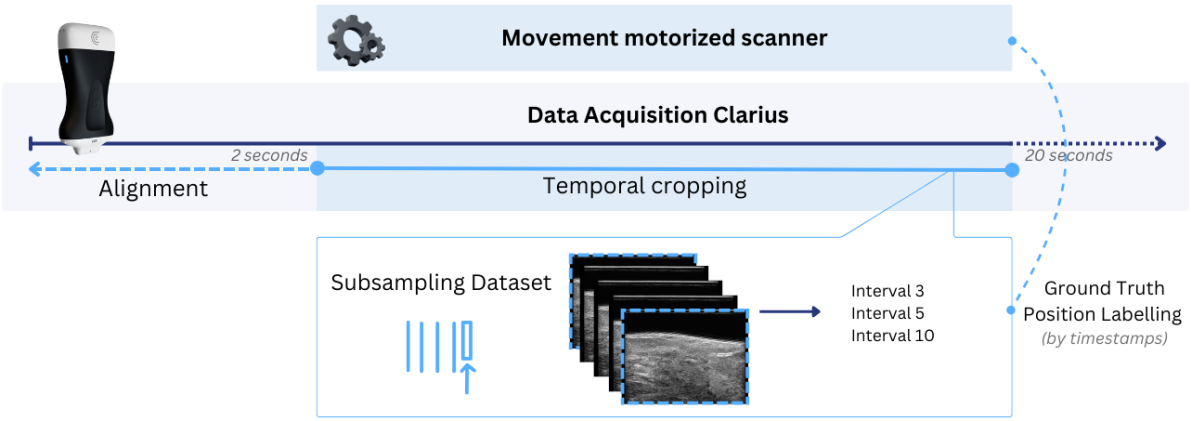


Figure 2.3: Workflow for preparing the collected data for data preprocessing, including alignment, temporal cropping, subsampling of the dataset and ground truth position labeling.

Temporal cropping

After initial alignment and synchronization of the US images, inertial data and positional data, the combined dataset was visualized to verify the accuracy of the alignment and to detect any potential lags or interruptions in data streaming due to network connectivity issues. Visualization involved plotting the positional data over time and overlaying it with the timestamps of the US images and IMU data. Segments with interruptions or inconsistencies between data sources were excluded to maintain the integrity of the dataset. To focus on the region of interest, the US images and inertial data were temporally cropped to include only relevant portions of the streamed data, corresponding to the movement interval of the motorized scanner. This step was essential to ensure that the datasets were correctly synchronized and included relevant data with no discrepancies that could adversely affect model training.

Subsampling of data

Due to the high frame rate of the US images and the relatively low scanning speed, the relative transformations between consecutive frames were minimal. However, translations between consecutive frames smaller than the pixel size of 0.04 mm by 0.04 mm, are indistinguishable in the image, resulting in a low signal-to-noise ratio (SNR) in the training data. Training the 3D reconstruction model on such data might cause it to focus on noise rather than meaningful movements, potentially degrading performance. To address this issue, additional datasets were created by subsampling the streamed data at lower frame rates, effectively increasing the displacement between frames and reducing the temporal resolution. By selecting every 4-th (interval of 3 frames), 6-th (interval of 5 frames) and 11-th frame (interval of 10 frames) as consecutive frame, three additional datasets were created and subjected to the data preprocessing workflow. However, it was essential to avoid excessively large frame intervals, as this could result in a loss of continuity and omission of important intermediate spatial information necessary for accurate 3D reconstruction. The characteristics of the subsampled datasets are presented in Table 2.2.

Table 2.2: Characteristics of subsampled datasets, representing the mean relative transformations between consecutive frames.

Subsampling Dataset	**Frames** (n) $(mean \pm sd)$	Δ**X** (mm) $(mean \pm sd)$	Δ**Y** (mm) $(mean \pm sd)$	Δ **Z** (mm) $(mean \pm sd)$	Δ**Tilt** (°) $(mean \pm sd)$
Consecutive	617 (±174)	0.018 (±0.006)	0.008 (±0.006)	0.080 (±0.017)	0.065 (±0.017)
Interval 3	154 (±43)	0.073 (±0.022)	0.031 (±0.025)	0.313 (±0.071)	0.254 (±0.064)
Interval 5	103 (±29)	0.109 (±0.033)	0.046 (±0.037)	0.469 (±0.106)	0.376 (±0.095)
Interval 10	56 (±15)	0.198 (±0.060)	0.080 (±0.066)	0.861 (±0.197)	0.663 (±0.167)

Ground-truth position labels

As a result of the alignment, the cropped and subsampled datasets were synchronized with the ground truth data from the motorized scanner through timestamps. As positional data from the motorized scanner were retrieved every 0.001 seconds, the nearest corresponding ground-truth position was identified based on the timestamp of each US image. The extracted machine position data were converted to transformation matrices including translation parameters (t_x, t_y, t_z) in millimeters and rotation parameters $(\theta_x, \theta_y, \theta_z)$ in radians, using the calibration values discussed in Section 2.1.2. In this setup, θ_y and θ_z, remained zero due to the limited DOF of the motorized scanner. As described in Section 2.1.3, tilting of the probe (θ_x) influenced the effective position of the imaging array in both z- and y-directions, which required a correction of the generated z-coordinates for motor input, to match the intended trajectory. Using the same trigonometric relationship, considering the offset of 82.5 mm between the probe's rotation point and its imaging array, the extracted motor position was now transformed to the location of the

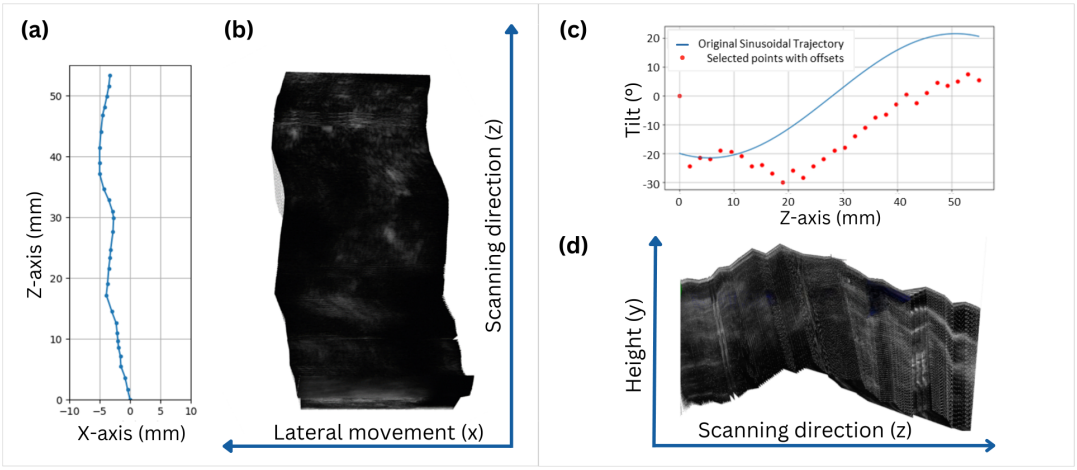


Figure 2.4: Illustration of a scanning trajectory on the horizontal an lateral planes. (a) A random generated grid (x, z). (b) Corresponding 3D US reconstruction using ground-truth US position labels (XZ-plane). (c) A random generated tilt grid (z, θ_x). (d) Corresponding 3D US reconstruction using ground-truth US position labels, illustrating the tilt and height along the scanning direction (ZY-plane).

imaging array. This provided a ground-truth trajectory of the imaging array, corresponding to the top border of the US images. An example of the US sweep using ground truth position labels, along with the generated trajectory grid, is presented in Figure 2.4. The final labels for the ground-truth trajectory of the images in each sequence were defined as:

- Local transformations $(\Delta t_x, \Delta t_y, \Delta t_z, \Delta\theta_x, \Delta\theta_y, \Delta\theta_z)$: Representing the relative translation and rotation parameters between consecutive frames.
- Global transformations $(t_x, t_y, t_z, \theta_x, \theta_y, \theta_z)$: Representing the absolute translation and rotation parameters for each frame, normalized to zero at the start position of the sweep. This provides the cumulative position and orientation of the probe at each time point.

Normalization of the labels was required to address differences in scales between parameters (millimeters vs. radians) and differences in the magnitude of movement between axes. Parameter-wise normalization was performed by dividing each parameter by its standard deviation computed from the training dataset, without subtracting the mean to maintain zero as the stationary value. The absence of motion was thereby accurately preserved. This approach ensured equal contribution of all parameters during model training, preventing those with larger magnitudes from dominating the loss function.

2.2.2. Preprocessing of input data

Following preparation of the sweep trajectory data, several preprocessing steps were applied to the US images and IMU datain order to enhance data quality and ensure consistency, facilitating effective training of the deep learning model. A summary of the conducted preprocessing steps per data type is presented in Figure 2.5.

Ultrasound images

The original US images of 480 × 640 pixels contained 22 black pixels on both the left and right sides due to the probe's field of view. As these pixels only serve as padding, the images were cropped to a multitude of 32 and resized to a square of 480 × 480 pixels, to meet the input requirements of the deep learning network. Pixel intensities, ranging between 0 and 255, were normalized to standardize the input data, using an approximation of the mean (μ) and standard deviation (σ) computed on the training set, centering the values around zero with unit variance.

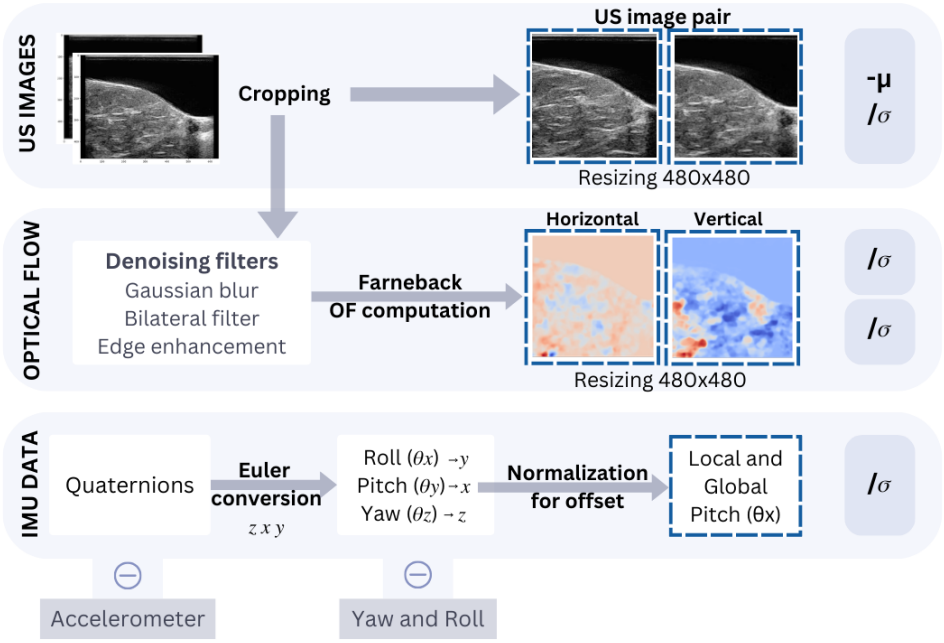


Figure 2.5: Overview of the preprocessing workflow, detailing steps for the US images, optical flow computation, and IMU data.

Optical flow

The ability of the model to capture in-plane motion between consecutive US frames, critical for accurate 3D reconstruction, is enhanced by including optical flow information as additional input for the model. Previous studies have demonstrated a consistently positive effect on performance using optical flow, providing a rationale for this approach [21–25]. Optical flow represents the apparent motion of anatomical structures, surfaces, and edges within a US image, as observed from changes in pixel values between consecutive frames [26]. In the context of 3D reconstruction, this can serve as valuable information, as the movement of pixels between successive US images occurs due to the movement of the probe. However, US images inherently contain a significant amount of speckle noise and artifacts, which can adversely affect the computation of optical flow by introducing false motion vectors. To improve the quality of the optical flow estimation, a series of denoising steps were applied to the US images before computation of the optical flow. The final parameters employed per denoising step are presented in Appendix B.

First, a Gaussian blur was applied which effectively suppresses noise by smoothing the image while preserving the overall structure. Subsequently, a bilateral filter was used, which considers both the spatial distance and intensity differences between pixels. This further reduces noise in homogeneous regions (small intensity differences) while maintaining sharp edges (high intensity differences), which are important for distinguishing structural features for motion estimation. Finally, a sharpening kernel was applied to accentuate edges and features within the image. This improves the detection of structural details within the US image, aiding in the accurate computation of optical flow vectors corresponding to actual movements.

After denoising, the Farnebäck optical flow algorithm was used to compute the optical flow between consecutive frames [27]. This algorithm estimates the displacement of each pixel between two images by modeling the motion as polynomial expansions. It includes several parameters that were fine-tuned based on experimental evaluations to optimize the balance between sensitivity to motion and robustness to noise. Specifically, to converge to optimal denoising filters and parameters for optical flow computation, test cases were created by manually translating frames by known amounts and di-

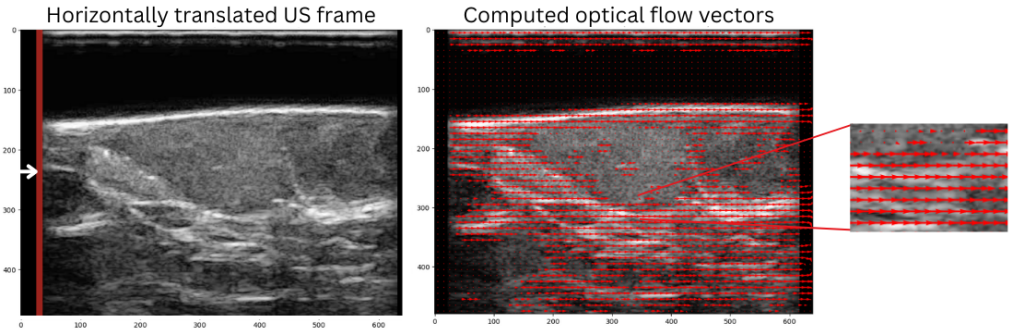


Figure 2.6: Visualization of a test case for a horizontally translated US frame, and its computed optical flow represented by motion vectors, downsampled for visualization purposes.

rections. The computed optical flow vectors were then compared to the expected movement to ensure that the optical flow algorithm and denoising steps accurately captured the true motion between frames and was not significantly affected by noise. An example of the computed optical flow, demonstrated over the original US image using the final configuration, is shown in Figure 2.6.

Finally, the optical flow vectors were normalized in the same manner as the transformation parameters using only the standard deviation computed across the training dataset. By not subtracting the mean, zero-valued vectors representing no movement between frames were preserved.

Inertial Measurement Unit (IMU)

The IMU integrated into the US probe provides valuable data comprising readings from the accelerometer, gyroscope, and magnetometer, as well as orientation quaternions. These data have the potential to offer information about the probe's orientation and translation during scanning. However, significant preprocessing is required to extract meaningful insights due to inherent noise and instability in the signals [28].

Despite the implementation of several preprocessing steps, significant challenges were encountered in obtaining reliable displacement estimates from the accelerometer data. The steps performed are described in Appendix C. As a result, the accelerometer data were excluded for further use in the model, focusing on the orientation data solely.

Processing orientation data

The orientation of the IMU is represented by quaternions, consisting of one real component (w) and three imaginary components (x, y, z). This four-dimensional number system provides a non-singular representation of 3D rotations. For the purpose of this study, the quaternions were converted to Euler angles to facilitate interpretation and alignment with the ground truth tilt angles of the motorized scanner. The conversion used the ZXY Euler sequence, which corresponds to rotations around the z, x, and y axes, respectively.

The global reference frame of the IMU sensor in the US probe, as discussed in Section 2.1.2, was aligned in a flat position. However, when mounted in the 3D-printed probe holder, the IMU sensor had an offset angle of 90 degrees in pitch, facilitating scanning in a vertical position, as shown in Figure 2.7. To account for this offset, the Euler angles were normalized by subtracting the mean of the initial stationary measurements acquired before the start of each sweep. This normalization set the initial orientation to zero degrees for all angles, ensuring alignment with the ground truth orientation labels at the start of the sweep.

To align the IMU data with the US images and motorized scanner coordinate system, the IMU axes were redefined accordingly (see Table 2.3). In this configuration, the pitch angle (θ_x) corresponds to

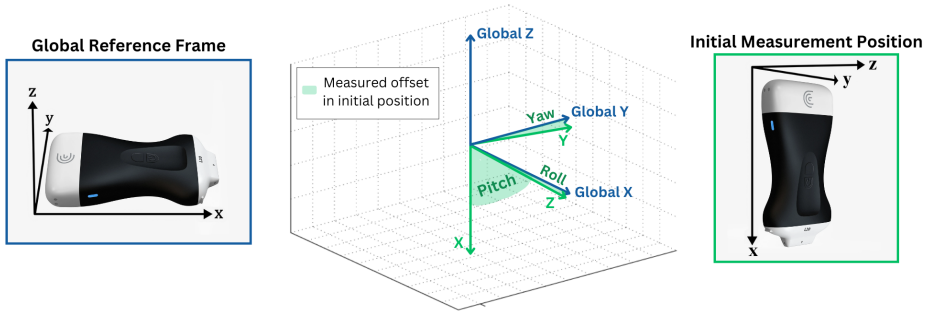


Figure 2.7: Example of the measured Euler angles yaw, pitch, and roll in initial measurement position, relative to the global reference frame, used for normalization to an initial offset of zero.

Table 2.3: Mapping of IMU Euler angles to the coordinate system of the image acquisition setup.

IMU Euler angle	Rotation axis IMU	Rotation axis setup	Motion
Yaw	Rotation around z-axis (θ_z)	z-axis (θ_z)	Scanning direction (stable)
Pitch	Rotation around y-axis (θ_y)	x-axis (θ_x)	Lateral (tilting)
Roll	Rotation around x-axis (θ_x)	y-axis (θ_y)	Height (stable)

the tilt motion introduced by the scanner, while yaw (θ_z) and roll (θ_y) angles are expected to remain constant, as the scanner only facilitates rotation around the x-axis. Any variation in yaw or roll would therefore indicate unintended movements or noise.

Exclusion of Yaw and Roll

During preliminary analyses, it was observed that the yaw angle (θ_z) exhibited significant instability and noise, primarily due to magnetic interference from external sources such as the stepper motors, the magnets stabilizing the specimen, or other surrounding metal structures. This interference affected the reliability of the magnetometer readings, which are critical for accurate yaw measurements. Given the constant yaw and roll angles during scanning, they were excluded from further analysis as this would merely introduce noise and did not contribute meaningful information. Focusing solely on the pitch angle (θ_x) provided the necessary information allowing for a more accurate estimation of the probe's orientation during scanning, critical for reconstructing the 3D position of the US images.

Local and Global pitch

To incorporate the IMU orientation data in the deep learning model, the local pitch and global pitch, were computed to match the ground-truth local and global rotation labels. For each US image, the closest IMU measurement in time was associated, providing the pitch angle (θ_x) at that time point. An example is illustrated in Figure 2.8. Both sets were normalized by dividing each parameter by its standard deviation, ensuring consistent scaling during model training and maintaining an angle of zero at zero.

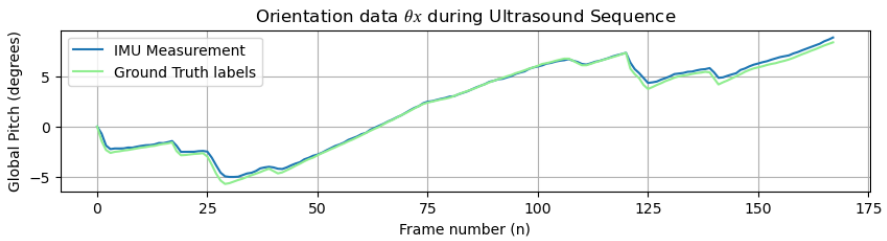


Figure 2.8: Global pitch (θ_x) of the IMU measurements compared to the ground truth orientation, over a US sequence of n frames.

2.3. Development of 3D reconstruction model

This section outlines the methodology employed in the development and training of the deep learning model, which has been designed for the purpose of 3D trajectory reconstruction of 2D US sequences. The section starts with a description of the baseline network architecture, followed by an overview of four categories of ablation experiments designed to enhance model performance. As illustrated in Figure 2.9, these include the integration of the IMU orientation, the implementation of generalization techniques, and training with different loss functions and different subsampled datasets (Section 2.3.2 – Section 2.3.5). Subsequently, the implementation details of model training are discussed, and the iterative methodology and outcome metrics used for evaluation of the ablation experiments are outlined, which ultimately lead to the final optimized model. Finally, the application of the 3D reconstruction model to the test set is presented, encompassing inference and visualization steps.

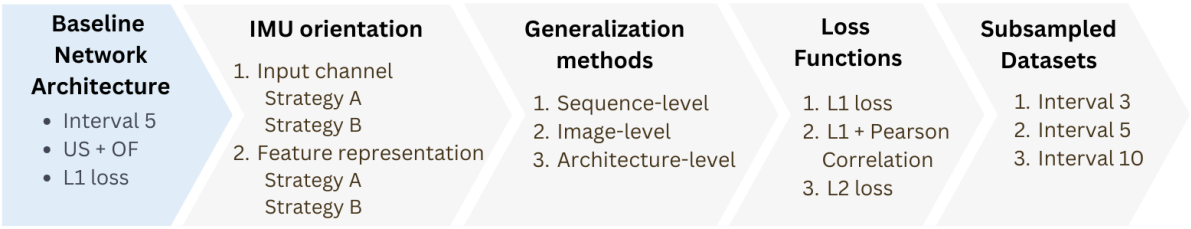


Figure 2.9: Flow diagram for the model development and iterative optimization of the baseline network architecture by ablation experiments 1 to 4.

2.3.1. Baseline Network Architecture

The proposed deep learning model combines a lightweight convolutional neural network (CNN) backbone with a Transformer decoder. By combining these two architectures, we leverage the CNN's ability to efficiently extract spatial features from individual images and the Transformer's capacity to model temporal dependencies across sequences [19]. Consequently, the model is capable of considering information from other frames within the sweep, thereby providing context that can enhance the prediction of transformations between frames. The objective of this design is to achieve an accurate 3D reconstruction while maintaining computational efficiency, which makes it suitable for real-time processing requirements.

The CNN backbone is based on the compact MobileNetV4 architecture, which has been pretrained on ImageNet [29]. This has been selected for its efficiency and suitability for handling the large amount of sequential data in this study. The CNN processes each successive pair of US images with corresponding optical flow data in order to extract high-level spatial features, resulting in a set of feature maps with reduced spatial dimensions and increased depth. To prepare these features for sequential modeling by the transformer, an embedding projection module transforms the feature maps into fixed-length embeddings. Positional encodings are added to the embeddings to incorporate temporal information, enabling the Transformer to distinguish the position of each frame within the sequence, a crucial aspect for modeling motion dynamics.

The Transformer decoder processes the entire sequence of embeddings to model temporal dependencies. It effectively captures motion dynamics and long-range relationships across the sequence without the limitations of recurrent neural networks, such as vanishing gradients or limited memory capacity that would be insufficient for the large number of frames in the sequence. It has a multi-layer, multi-head self-attention mechanism, which allows the model to weigh the relevance of different frames when making predictions.

Finally, the enriched embeddings from the Transformer are fed into four distinct classification heads

to predict the transformation parameters: local translation (LT), local rotation (LR), global translation (GT), and global rotation (GR). By using separate classification heads for local and global transformations, the model can discern both immediate changes between frames and movements relative to the initial position, thereby facilitating specialized learning leading to potentially more accurate predictions.

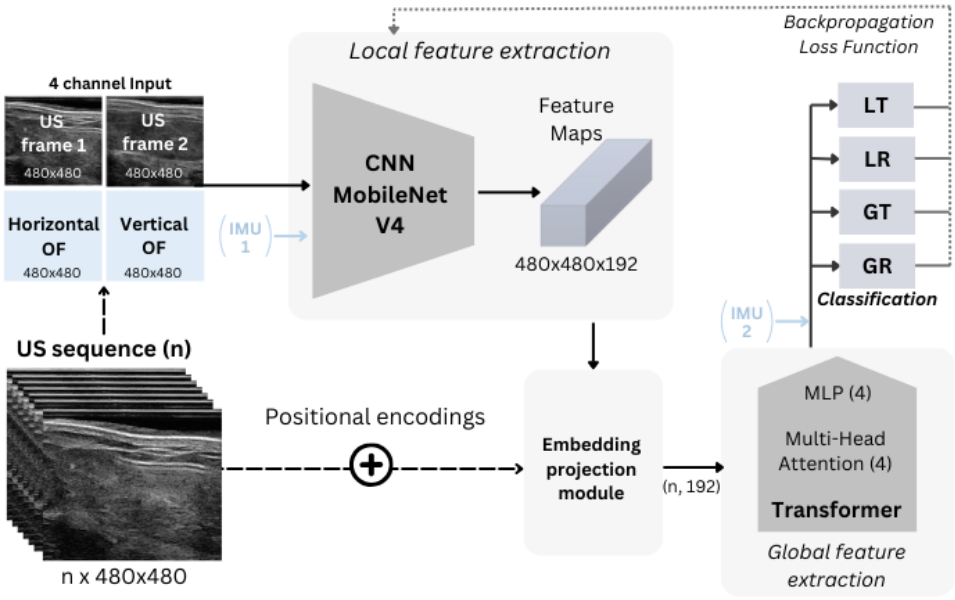


Figure 2.10: Overview of the CNN-Transformer architecture used as baseline model. In addition, it incorporates the IMU 1 and IMU 2 components, which pertain to the integration methods outlined Section 2.3.2.

2.3.2. Experiment 1: Integration of IMU orientation

To enhance the model performance for 3D reconstruction, the baseline network architecture was augmented with the integration of IMU orientation data. Given the lack of consensus on an optimal methodology for incorporating IMU data, two integration strategies, illustrated in Figure 2.10, were examined to assess effectiveness in this context,.

In the first strategy, the local and global pitch angles derived from the IMU were employed as supplementary input channels, in conjunction with the US images and optical flow data. By directly incorporating the IMU orientation data into the model's input, the network has the potential to learn complex relationships between the IMU measurements and spatial features extracted from the images. This approach allows the model to leverage spatial correlations by providing explicit orientation information at the earliest stage of processing.

The second strategy involved concatenating the IMU orientation data to the network's feature representation prior to the classification heads. By involving the IMU data at this stage, the model retains the distinct features learned from the images and optical flow while integrating the orientation information to inform the final predictions. This method reduces the risk of overfitting or over-engineering of features, as the IMU data are introduced after the main feature extraction and temporal modeling processes, allowing for a clearer contribution of the IMU to the final output.

For both methods, referred to as 1A and 2A, the models were trained to predict both orientation and translation labels, utilizing the full potential of the combined data sources. In addition, a second evaluation strategy was examined for both models, referred to as 1B and 2B, whereby the translation parameters were predicted by the model, while the orientation labels adhered strictly to the local and global pitch angles provided by the IMU. This resulted in four final methods being evaluated for the integration with IMU orientation.

2.3.3. Experiment 2: Generalization methods

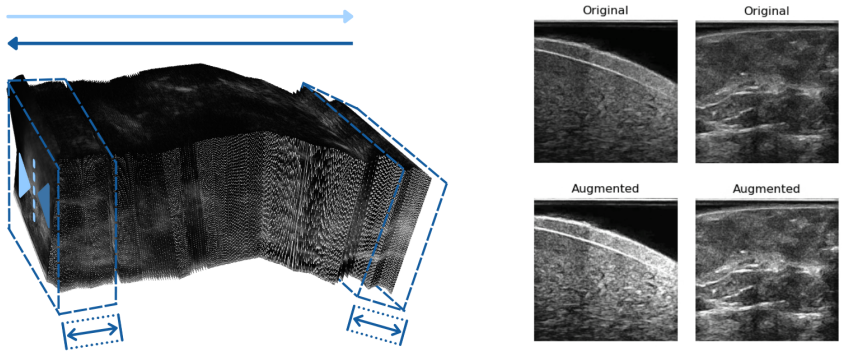
In the second experiment to improve model performance, data augmentation strategies and architectural modifications were employed to prevent overfitting and improve the model's generalization capabilities. Data augmentation was applied at two different levels; sequence-level and image-level. This was done with the objective of introducing variability and better representing the diversity of trajectories and images. As a third method, architectural modifications were introduced, applying regularization directly within the network's architecture. This resulted in three distinct models, each evaluated for their impact on overall performance.

Data augmentation

At the level of sequence augmentation, three techniques were applied; flipping the temporal order of frames within the sequence, mirroring the images horizontally and random cropping by using different start and end points within the sequence. These augmentations were carefully designed to maintain the spatial relationships between frames, with corresponding optical flow and transformation labels adjusted accordingly to reflect the augmentations applied. This contributes to the model's robustness with regard to variations in temporal order, spatial characteristics, as well as sequence length.

Data augmentation at image level included random applied adjustments to the image including brightness, contrast and gamma correction, within the limits of preset magnitudes. These adjustments introduced variability in the appearance of the US images, enhancing the model's robustness to variations of imaging conditions or settings.

According to the augmentation method, whether sequence-level or image-level, the augmentation technique was applied to all images in the sequence with a probability of 0.5, thereby enabling multiple augmentations to occur concurrently. A visual representation of the three sequence-level augmentation techniques is presented in Figure 2.11(a), and an example of image-level augmentation in Figure 2.11(b).



(a) Sequence Augmentation **(b) Image Augmentation**

Figure 2.11: (a) Illustration of sequence augmentation techniques, including direction change, random cropping or horizontal mirroring. (b) Example of image augmentation by adjustments in brightness, contrast and gamma correction of two US images.

Network architecture level

Regularization at the level of the network architecture included deployment of dropout layers and input channel dropout. Dropout layers were integrated in the embedding and classification layers, with dropout rates set to 0.2 and 0.3, respectively. During training, dropout randomly deactivates a fraction of neurons to prevent certain neurons from relying too heavily on specific patterns and encouraging the network to learn more robust and generalizable features. During inference all neurons remain active,

and their outputs are scaled to reflect the training dropout rates, ensuring consistency in predictions. In addition, an input channel masking method was applied, where one of the input channels, either the first US frame, the second US frame, or the horizontal or vertical components of the optical flow, was randomly set to zero with equal probability during training. This approach forces the model to rely on multiple modalities and prevents over-reliance on a single input channel, enhancing its ability to generalize.

2.3.4. Experiment 3: Loss functions

As third experiment, alternative loss functions were employed in model training. The adopted network architecture is capable of predicting four distinct types of transformations through its four classification heads, namely local translation, local rotation, global translation, and global rotation. In order to train the model effectively, separate loss functions are computed for each of these outputs, focusing on both translation and rotation parameters for both local and global parameters. The total loss used for backpropagation is the sum of the individual losses computed for each transformation type. This approach allows the model to learn multiple related tasks simultaneously, while ensuring that each task contributes appropriately to the learning process after normalization.

Specifically, three different types of loss functions were employed to evaluate and optimize the model's performance:

Mean Absolute Error (MAE)

The initial approach, employed in the baseline model, entailed the Mean Absolute Error (MAE), also known as the L1 loss. This loss function is robust to outliers and provides a stable gradient, ensuring that the model minimizes the average absolute difference between the predicted and ground truth parameters over all frames in a sequence. It is defined as:

$$\text{MAE} = \frac{1}{N} \sum_{k=1}^{N} \left\| T_k^{\text{pred}} - T_k^{\text{true}} \right\|_1 \tag{2.1}$$

where:

- N is the number of frames.
- T_k^{pred} and T_k^{true} are the predicted and ground truth transformation matrices, consisting of either translation (t_x, t_y, t_z) or rotation $(\theta_x, \theta_y, \theta_z)$ parameters for frame k, representing local or global coordinates.
- $\|\cdot\|_1$ denotes the L1 norm, which is the sum of absolute differences.

However, the MAE treats each frame independently and does not account for temporal dependencies, potentially leading to drift accumulation over time.

MAE and Pearson Correlation Loss

To address the limitations of using MAE alone, a combined loss function was introduced, inspired by approaches in literature that incorporate a Pearson Correlation Loss term [21, 30–32]. The MAE component ensures that the model minimizes the average absolute error between the predicted and ground truth parameters, focusing on the frame-wise magnitude of errors. The Pearson Correlation Loss complements this by focusing on the overall shape and trends of the predicted sequences by measuring the linear relationship between predicted and ground-truth parameters. It ensures that the predicted temporal dynamics align with those of the ground truth. By combining these two loss functions, the model could benefit from both point-wise accuracy through the MAE, and temporal consistency through the

correlation loss. The combined loss function is defined as:

$$L_{\text{total}} = L_{\text{MAE}} + \left(1 - \frac{\text{Cov}\left(T^{\text{pred}},\ T^{\text{true}}\right)}{\sigma_{T^{\text{pred}}}\sigma_{T^{\text{true}}}}\right) \tag{2.2}$$

where:

- L_{MAE} is the MAE loss computed as in Equation (2.1).
- $\text{Cov}\left(T^{\text{pred}},\ T^{\text{GT}}\right)$ is the covariance between the predicted and ground truth sequences.
- $\sigma_{T^{\text{pred}}}$ and $\sigma_{T^{\text{true}}}$ are the standard deviations of the predicted and ground truth sequences.
- $\frac{\text{Cov}(T^{\text{pred}},\ T^{\text{true}})}{\sigma_{T^{\text{pred}}}\sigma_{T^{\text{true}}}}$ is the Pearson correlation coefficient.

By subtracting the Pearson correlation coefficient from 1, the loss function encourages higher correlation (lower loss) between the predicted and ground truth parameters, encouraging the model to capture the correct temporal trends and relationships.

Mean Squared Error (MSE)

As a third loss function, the MSE, or L2 loss, was employed. It calculates the average squared difference between the predicted and ground truth parameters, and is defined as:

$$\text{MSE} = \frac{1}{N}\sum_{k=1}^{N}\left\|T_k^{\text{pred}} - T_k^{\text{true}}\right\|_2^2 \tag{2.3}$$

where $\|\cdot\|_2$ denotes the L2 norm. The MSE penalizes larger errors more heavily due to the squaring of the error terms. This property makes the MSE more sensitive to outliers and noise, compared to the MAE. However, it encourages the model to focus on minimizing significant deviations, which is valuable when large errors are particularly undesirable.

2.3.5. Experiment 4: Subsampled datasets

As a final experiment for model optimization, the model was trained on the three subsampled datasets described in Section 2.2.1. Up until this point, the dataset with an interval of 5 frames was utilized. Subsequently, training was conducted on datasets with reducing interval of 3 frames, with smaller interframe distances, and an increasing interval of 10 frames, with larger interframe distances. This resulted in a different distribution of transformation parameters, requiring recomputation of normalization parameters. As the average number of frames in a sequence varied per dataset, the dimensions for random cropping in data augmentation were adjusted.

2.3.6. Implementation details

In all experiments, except for the specific modifications examined, consistent training parameters and procedures were maintained to ensure that any observed differences in performance could be attributed solely to the experimental variables.

For model development, the complete dataset, comprising both phantom and ex-vivo data, was randomly divided into a training set (80%) and a test set (20%) without patient stratification, as each trajectory was unique and no overlap between sweeps occurred. The test set was exclusively reserved for the final evaluation and visualization. To evaluate the effect of the different experiments, a 5-fold cross-validation strategy was employed for the training set, with a fixed random seed to ensure consistent data splits across all experiments. This facilitated reliable comparisons between different models and configurations.

The CNN and Transformer decoder were implemented using PyTorch (version 2.3.1) and trained on a single GPU. The input of the CNN included pairs of US images and their corresponding optical flow components (in horizontal and vertical direction), resulting in four input channels per sample. Training was performed with a batch size of 1, as transformers require input sequences within a batch to be of equal length. Due to the variable lengths of US sequences, a larger batch size would necessitate padding all sequences to match the longest one, leading to computational inefficiency. Models were optimized using different loss functions described in Section 2.3.4, with an Adam optimizer and an initial learning rate of 5×10^{-4}. A one-cycle learning rate scheduler was used, which briefly increases the learning rate to a peak before gradually decreasing it over the course of training. This approach helps in achieving better convergence and avoiding local minima. Models were trained for 100 epochs, providing sufficient time for convergence without significant overfitting. Model performance was monitored using the validation loss and evaluation metrics on the cross-validation folds. Consistent decreases in validation loss and stable evaluation metrics indicated proper convergence.

2.3.7. Ablation experiments

Ablation experiments were conducted to systematically assess the impact of specific model components on the performance, thereby elucidating the contribution of individual elements within the model architecture.

The experiments were performed iteratively, with the best model from each experiment serving as the new baseline for the subsequent one. This allowed for a systematic assessment of each modification while limiting the number of models trained. The sequence of experiments was designed to proceed from those expected to have the greatest impact on performance to those with potentially lesser effects. This approach began with the integration of IMU data, as prior studies indicate its significant influence on model performance [21, 22, 30, 31]. Subsequent experiments focused on methods to enhance generalization and robustness, such as data augmentation and dropout. Finally, adjustments like alternative loss functions and dataset subsampling were evaluated.

Within each experiment, model performance was assessed per parameter using the outcome metrics discussed in Section 2.3.8. Models were ranked based on these metrics, and the best-performing model was statistically compared to the baseline model per parameter. As the outcome metrics were not normally distributed, a Wilcoxon signed-rank test was employed to assess significant improvements [33]. A p-value less than 0.05, after applying Holm's correction for multiple testing due to the number of outcome measures and parameters, was considered significant [34]. A model was considered to perform better than the baseline if it demonstrated significant improvements in more parameters than those exhibiting a decline in performance. This model subsequently served as the new baseline for the next experiment and culminated in the configuration of the final model.

2.3.8. Quantitative outcome measures

To evaluate the performance of the models, several quantitative outcome measures were employed, focusing on the accuracy of the predicted transformation parameters and the reconstruction of the probe's trajectory.

Mean Absolute Error (MAE)

The metric employed for the assessment of error for each distinct parameter along each axis, at each time step, was the MAE. It quantifies the mean of the absolute difference between the predicted and ground-truth transformation parameters. This metric directly corresponds to the MAE loss function defined in Equation 2.1.

Final Drift (FD) and Final Drift Rate (FDR)

Another key metric in trajectory reconstruction is the final drift (FD), defined by the Euclidean distance (in millimeters) between the positions of the center of the last frame of the estimated trajectory and the last frame of the ground truth trajectory. Since the prediction of the trajectory of the sweep is based on a transformation between successive frames, errors accumulate over the length of the sweep. This accumulated error is reflected in the metric FD.

To account for the varying lengths of different sweeps, the Final Drift Rate (FDR) was computed by normalizing the FD by the total length of the ground truth trajectory, and is expressed as a percentage. The FDR provides a relative measure of the drift, making it comparable across sequences of different lengths. It is important to note that a minimal FD or FDR does not necessarily indicate a satisfactory reconstruction for complex scan strategies, as only the location of the last frame is taken into account.

Selection for evaluation of ablation experiments

To fairly assess significant improvements of a model compared to the baseline, a selection of transformation parameters and outcome metrics was made. Since the rotations around the y-axis (θ_y) and z-axis (θz) remained constant at zero in the dataset, the evaluation of orientation parameters focused on rotation around the x-axis (θ_x). Together with three translation parameters, this resulted in a selection of four transformation parameters assessed using the MAE.

Additionally, as similar data was used in all ablation experiments (Section 2.3.2 - Section 2.3.5), thus had consistent lengths, the FD and FDR provided similar information. While FD is a valuable metric in assessing performance, it was redundant in the context of ablation experiments. Therefore, the FDR was selected for evaluation of the experiments, allowing for better comparison across different sequences as it normalizes for the length of the sweep.

For the first three ablation experiments, the MAE and FDR were considered for both local and global transformations, resulting in a total of ten metrics for evaluation. In the fourth ablation experiment, utilizing different subsampled datasets, the focus was on solely the global parameters assessed by MAE and FDR. The varying data distributions, specifically the interframe distances, impacted the values of local MAE, making comparisons challenging. However, global transformations provided a consistent basis for evaluation. An overview of the selected parameters and metrics per ablation experiment is provided in Table 2.4.

Table 2.4: Selected metrics for evaluation of the ablation experiments

Ablation Experiment	Metrics (local parameters)	Metrics (global parameters)
1. IMU	MAE $(t_x, t_y, t_z, \theta_x)$	MAE $(t_x, t_y, t_z, \theta_x)$
2. Generalization	FDR	FDR
3. Loss functions		
4. Subsampling datasets	N/A	MAE $(t_x, t_y, t_z, \theta_x)$
	N/A	FDR

Selection for evaluation of the final model

For final model evaluation, the FD, FDR and MAE excluding rotations around the y-axis (θ_y) and z-axis (θ_z), were considered for both local and global transformations.

The CNN-Transformer architecture further provides the flexibility to predict two distinct output types: local and global parameters. As both local and global parameters can be used for 3D reconstruction, this allows for a comparative analysis. Specifically, the local labels, representing frame-to-frame transformations, were accumulated over time to form global labels. These cumulative local labels were

then compared to the global labels predicted directly by the Transformer using the MAE and FDR metrics. Statistical comparisons were conducted using the Wilcoxon signed-rank test due to the non-normal distribution of the data, with Holm's correction applied to adjust for multiple comparisons.

2.3.9. Inference and visualization of 3D reconstruction model

While the quantitative evaluation of 3D trajectory reconstructions provides a basis for selecting the most accurate model based on trajectory errors, visualizing the 3D reconstructions offers valuable insights into the practical applicability of the model. Accordingly, following the quantitative assessment of the final optimized model on the separately reserved test set, the inference and visualization of a sample within the test set was demonstrated. The predicted transformations were reconstructed in 3D as follows:

1. Positioning frames in space: The predicted transformations were applied to position each US frame in 3D space, reconstructing the trajectory of the probe. This spatial arrangement provides a visual representation of the sweep and demonstrates the model's ability to accurately track the probe's movement, aligning with the scope of this thesis.

2. Filling voxel gaps: Nearest neighbour interpolation was employed to fill the voxel spaces between frames. This step aimed to generate a continuous 3D structure, enhancing the visual completeness of the reconstruction and to facilitate the interpretation of the anatomical features.

3. Segmentation: As a practical use case, anatomical structures in the ex-vivo specimen were manually segmented on the 2D US frames and reconstructed to a 3D volume using nearest neighbor interpolation.

3

Results

3.1. Data acquisition

3.1.1. Dataset characteristics

The dataset acquired in this thesis comprised a total of 361 US sweeps, including 340 sweeps derived from ex-vivo specimens and 21 sweeps from a phantom model. Multiple sweeps were acquired per specimen, each at different locations, to capture a diverse range of tissue characteristics ensuring no overlap. An overview of the dataset, categorized by tissue type, is presented in Table 3.1.

Each sweep captured a unique trajectory to encompass variability in probe movement. The trajectory characteristics extracted from the ground truth data across the entire dataset revealed a median sweep length of 49.97 mm (range: 9.9-60.2 mm). The lateral movements along the x-axis ranged from −5 mm to 7.8 mm, while vertical movements along the y-axis ranged from 0 to 15.9 mm, relative to the start position. The probe orientation varied by tilt movements ranging from -36.2° to 16.6°. Lastly, motion speed of the US probe varied between 1.8 mm/s and 2.6 mm/s.

Table 3.1: Overview of acquired dataset categorized by tissue type.

Data type	**Tissue**	**Number of inclusions**	**Number of US sequences**
Ex-vivo	Breast	16	244
	Colorectal	5	80
	Sarcoma	1	16
Phantom	IOUSFAN	1	21
Total		**23**	**361**

3.2. Development of 3D reconstruction model

This section presents the results of the developed baseline network architecture utilizing US images and corresponding optical flow data, followed by the results of the ablation experiments. The quantitative outcomes of each experiment, obtained through cross-validation, and key findings are detailed in Section 3.2.1. An iterative approach was used to identify the best-performing model within an experiment. Whether significant improvements were achieved compared to the baseline model, was

assessed using a Wilcoxon signed-rank test (p-value < 0.05). This iterative evaluation process led to the selection of the final model, which was evaluated on the reserved test set in Section 3.2.2. Lastly, the final model was demonstrated for inference through a visual representation of the 3D reconstruction in Section 3.2.3.

3.2.1. Quantitative outcomes of ablation experiments

Experiment 1: Integration of IMU orientation

The first ablation experiment examined the added value of integrating IMU orientation data by two methodologies; using it as additional input channel alongside the US images and optical flow (IMU 1), and concatenation of the IMU orientation with the feature representation prior to classification (IMU 2). Each method was evaluated using two distinct strategies, designated as A and B. In strategy A, the evaluated transformation parameters were predicted by the model, whereas in B, the orientation labels were replaced with the IMU orientation data before quantitative evaluation.

Quantitative outcome metrics for strategy A are represented in Table 3.2, comparing the baseline, IMU 1A, and IMU 2A models. With regard to the local parameters, model IMU 2A showed an improved MAE for t_y and θ_x in comparison to both the baseline model and IMU 1A. Specifically, IMU 2A achieved a significant reduction in MAE for t_y from 0.018 ± 0.019 mm (baseline) to 0.012 ± 0.008 mm, while t_x showed a significant increase in MAE for IMU 2A from 0.044 ± 0.028 mm (baseline) to 0.052 ± 0.023 mm. At the global level, both IMU 1A and 2A demonstrated significant improvements over the baseline for all transformation parameters, with a particular note on the MAE of t_z decreasing from 6.606 ± 21.340 mm in baseline to 2.684 ± 1.923 mm in IMU 2A.

In strategy B, replacing the predicted orientation with IMU data affected only θ_x and did not influence the FDR due to the relatively small magnitude of θ_x. Accordingly, Figure 3.1 illustrates the MAE for the rotation parameter θ_x to evaluate both models and strategies. Integration of IMU data consistently improved the MAE for θ_x in comparison to the baseline, following a similar trend for both local and global labels. IMU 2A outperformed other models significantly, even the models using strategy B, with a local MAE of $0.071 \pm 0.031°$ and global MAE of $0.393 \pm 0.295°$, compared to $0.098 \pm 0.071°$ and $1.342 \pm 1.418°$ for baseline, respectively. This suggests that integrating IMU data within the feature representation (IMU 2A) is more effective than simply replacing the output during evaluation.

In conclusion, IMU 2A demonstrated the best overall performance among the evaluated methods. Despite a significant increase in MAE for t_x at the local level, IMU 2A showed significant improvements in six out of ten evaluated metrics compared to the baseline. Therefore, IMU 2A was considered superior and was selected as the new baseline for subsequent experiments.

Table 3.2: Quantitative outcome metrics for the baseline, IMU 1A, integrating the orientation as input channel, and IMU 2A integrating the orientation prior to classification, with the best performing method presented in bold.

Model	**MAE** $(mean \pm sd)$				**FDR** $(mean \pm sd)$
Local Transformations	t_x (mm)	t_y (mm)	t_z (mm)	θ_x (°)	(%)
Baseline (US + OF)	0.044 ± 0.028	0.018 ± 0.019	0.103 ± 0.069	0.098 ± 0.071	10.46 ± 7.81
IMU 1A	0.057 ± 0.022	0.018 ± 0.020	0.127 ± 0.063	0.093 ± 0.049	16.37 ± 17.93
IMU 2A	0.052 ± 0.023*	0.012 ± 0.008*	0.110 ± 0.053	0.071 ± 0.031*	11.13 ± 10.48
Global Transformations	t_x (mm)	t_y (mm)	t_z (mm)	θ_x (°)	(%)
Baseline (US + OF)	1.430 ± 1.633	0.211 ± 0.426	6.606 ± 21.340	1.342 ± 1.418	14.48 ± 12.40
IMU 1A	1.142 ± 0.686	0.163 ± 0.340	3.018 ± 2.095	0.815 ± 0.741	14.18 ± 13.88
IMU 2A	1.068 ± 0.628*	0.082 ± 0.122*	2.684 ± 1.923*	0.393 ± 0.295*	13.83 ± 13.35*

* indicates a significant difference with the baseline model, with a p-value < 0.05 after Holm's correction.

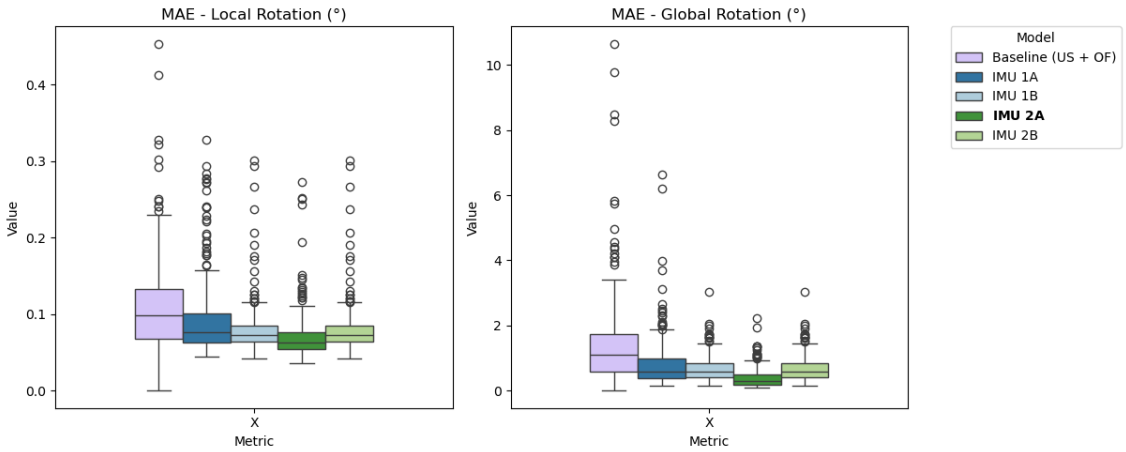


Figure 3.1: Comparison of the MAE (°) of the local and global rotation (θ_x) for model IMU 1A, IMU 1B, IMU 2A and IMU 2B.

Experiment 2: Generalization methods

The second ablation experiment evaluated the impact of three generalization methods, including data augmentation at sequence level, data augmentation at image level, and generalization methods at the level of network architecture, which included dropout layers and input channel masking.

The quantitative outcomes for these models are represented in Table 3.3. The model employing data augmentation at sequence-level demonstrated the lowest errors across the majority of metrics and was the only model showing improvements compared to the baseline model (IMU 2A). In contrast, both the image-level augmentation and architecture-level modifications performed comparable to those of the baseline model or exhibited minor reductions in performance.

Specifically, sequence-level augmentation exhibited a significant reduction in the MAE of local transformation parameters t_x, t_y, t_z, θ_x compared to the baseline. At the global level, significant improvements were noted in t_x, t_z. The FDR significantly improved for local parameters from 11.13 ± 10.48% to 9.11 ± 10.05%, and global parameters from 13.83 ± 13.36% to 11.47 ± 11.28%, respectively. Although differences were not as extensive as those observed in employing IMU integration, sequence-level augmentation effectively enhanced the model's performance by introducing more variation in sweep trajectories. Based on these findings, the sequence-level augmentation model was deemed to outperform the baseline and was selected as the new baseline model for subsequent experiments.

Table 3.3: Quantitative outcome metrics for the generalization methods, employed at sequence-level, image-level and architecture-level, compared to the baseline (IMU 2A), with the best performing method presented in bold.

Model	**MAE** $(mean \pm sd)$				**FDR** $(mean \pm sd)$
Local Transformations	t_x (mm)	t_y (mm)	t_z (mm)	θ_x (°)	(%)
Baseline (IMU 2A)	0.052 ± 0.023	0.012 ± 0.008	0.110 ± 0.053	0.071 ± 0.031	11.13 ± 10.48
Sequence-level	0.044 ± 0.022*	0.011 ± 0.008*	0.102 ± 0.068*	0.067 ± 0.032*	9.11 ± 10.05*
Image-level	0.053 ± 0.023	0.012 ± 0.008	0.114 ± 0.053	0.074 ± 0.033	11.75 ± 11.88
Architecture-level	0.055 ± 0.024	0.014 ± 0.011	0.117 ± 0.083	0.083 ± 0.074	12.50 ± 11.20
Global Transformations	t_x (mm)	t_y (mm)	t_z (mm)	θ_x (°)	(%)
Baseline (IMU 2A)	1.068 ± 0.628	0.082 ± 0.122	2.684 ± 1.923	0.393 ± 0.295	13.83 ± 13.36
Sequence-level	0.861 ± 0.593*	0.088 ± 0.169	2.352 ± 1.697*	0.412 ± 0.321	11.47 ± 11.28*
Image-level	1.051 ± 0.624	0.088 ± 0.182	2.953 ± 2.293	0.392 ± 0.338	14.37 ± 13.58
Architecture-level	1.201 ± 0.732	0.123 ± 0.241	2.878 ± 2.052	0.608 ± 0.946	14.01± 11.88

* indicates a significant difference with the baseline model, with a p-value < 0.05 after Holm's correction.

Experiment 3: Loss functions

The effect of different loss functions was evaluated in the third ablation experiment, encompassing the L1 loss within the baseline model, a combined loss incorporating the L1 loss with a Pearson correlation term, and the L2 loss. A comparison of the selected outcome metrics for these models yielded no statistically significant differences, as presented in Table 3.4. The model utilizing the L2 loss demonstrated slight, but statistically insignificant, improvements in several metrics. However, in evaluation beyond the selected metrics, looking at the final drift, the L2 loss demonstrated a significant reduction from 4.749 ± 3.691 mm to 4.494 ± 3.629 mm. Accordingly, the L2 loss was selected, while acknowledging that the L1 and combined loss are equally viable alternatives, given the minimal difference and high standard deviations observed.

Table 3.4: Quantitative outcome metrics for the baseline model employing the L1 loss, the combined loss (L1 + Pearson correlation), and the L2 loss, with the best performing method presented in bold.

Model		**MAE** $(mean \pm sd)$			**FDR** $(mean \pm sd)$
Local Transformations	t_x (mm)	t_y (mm)	t_z (mm)	θ_x (°)	(%)
Baseline (L1)	0.044 ± 0.022	0.011 ± 0.008	0.102 ± 0.068	0.067 ± 0.032	9.114 ± 10.047
L1+Pearson	0.045 ± 0.023	0.011 ± 0.007	0.101 ± 0.056	0.067 ± 0.033	9.009 ± 8.126
L2	0.044 ± 0.023	0.011 ± 0.007	0.100 ± 0.052	0.066 ± 0.029	8.995 ± 8.575
Global Transformations	t_x (mm)	t_y (mm)	t_z (mm)	θ_x (°)	(%)
Baseline (L1)	0.861 ± 0.593	0.088 ± 0.169	2.352 ± 1.697	0.412 ± 0.321	11.466 ± 11.284
L1+Pearson	0.877 ± 0.634	0.084 ± 0.097	2.339 ± 1.820	0.413 ± 0.290	11.446 ± 11.250
L2	0.863 ± 0.589	0.082 ± 0.102	2.336 ± 1.852	0.394 ± 0.270	10.943 ± 10.851

* indicates a significant difference with the baseline model, with a p-value < 0.05 after Holm's correction.

Experiment 4: Dataset

Lastly, different subsampled datasets were utilised, in which the evaluation focused on global parameters only. An example illustrating the effect of subsampling of the dataset on the 3D US reconstruction is shown in Figure **??**. As presented in Table 3.5, subsampling with a smaller interval (interval 3) than the baseline (interval 5) did not result in enhanced quantitative outcomes. Instead, the outcomes were comparable, with only one instance of a significant difference, namely an increased FDR from 10.94 ± 10.85% to 11.54 ± 9.81%. With regard to subsampling with a larger interval, worse performance was seen, indicating that larger interframe distances can result in a loss of information. Consequently, the baseline model utilizing a dataset subsampled with an interval of 5 was maintained.

Table 3.5: Quantitative outcome metrics for training with different subsampled datasets, with the best performing method presented in bold.

Model		**MAE** $(mean \pm sd)$			**FDR** $(mean \pm sd)$
Global Transformations	t_x (mm)	t_y (mm)	t_z (mm)	θ_x (°)	(%)
Baseline (Interval 5)	0.863 ± 0.589	0.082 ± 0.102	2.336 ± 1.852	0.394 ± 0.270	10.94 ± 10.85
Interval 3	0.809 ± 0.572	0.082 ± 0.119	2.426 ± 1.897	0.398 ± 0.282	11.54 ± 9.83
Interval 10	1.001 ± 0.632	0.084 ± 0.109	2.621 ± 1.806	0.426 ± 0.288	12.00 ± 11.19

* indicates a significant difference with the baseline model, with a p-value < 0.05 after Holm's correction.

Summary of ablation experiments

In conclusion, the ablation experiments demonstrated significant improvements over the baseline model. Incorporating US frames, corresponding optical flow, and concatenated IMU orientation into the feature representation prior to the classification layer resulted in enhanced performance. Additionally, employment of sequence-level augmentation and utilizing the L2 loss function further improved the model's accuracy. Figure 3.2 illustrates the predicted global trajectories of these optimized models compared to the baseline and ground truth trajectories for two different sequences. The predicted trajectory of the final model is clearly less noisy and more similar to the ground truth model than the previous model versions. In Appendix D, a similar representation of the reconstructed trajectories is shown using the predictions of the local parameters. Although it shows a similar overall trend in optimization compared with the global parameters, it is noteworthy that the trajectories using local parameters appear slightly smoother, especially focusing on the previous model versions.

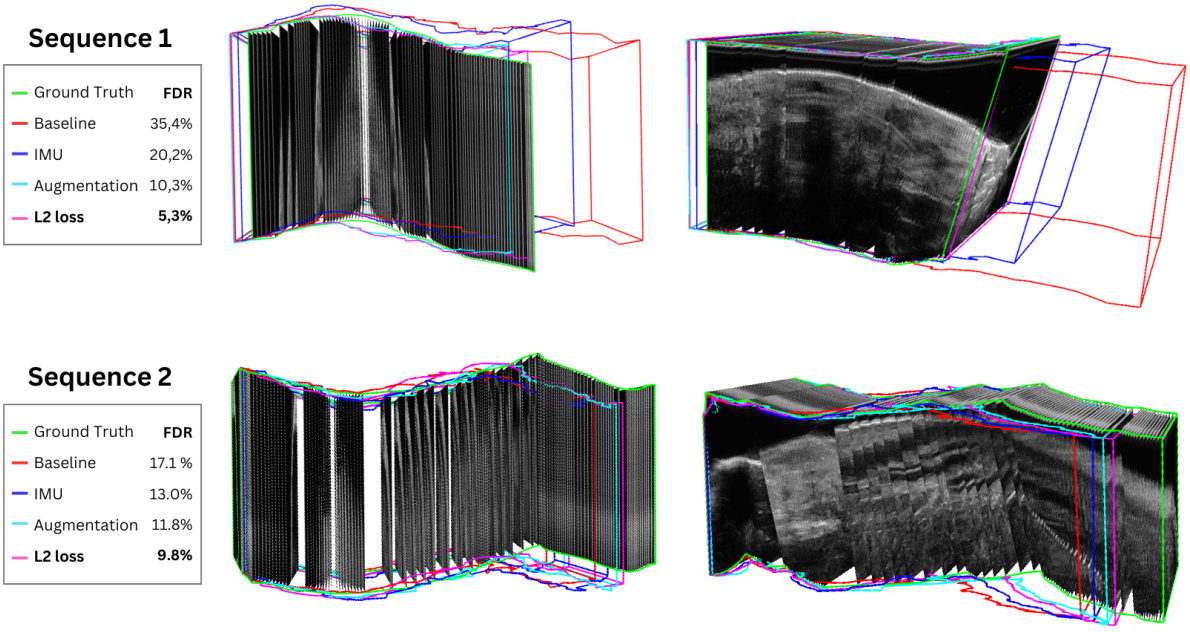


Figure 3.2: Comparison of the predicted trajectories from the selected models against the ground truth trajectory, based on the global parameters, including top (left) and side views (right) of two sample sweeps.

3.2.2. Quantitative outcomes of the final model

The final optimized model, incorporating the enhancements identified during the ablation experiments by utilizing five-fold cross-validation, was trained on the complete training set. The training process required 4 hours and 22 minutes. The quantitative results of the final model evaluated on the reserved test set are presented in Table 3.6.

Comparative analysis of the local and global parameters

The global parameters obtained by accumulating the local parameters (denoted as $\sum$ Local) were compared with the global labels predicted directly by the network (denoted as global). Statistical analysis revealed that the parameters t_y and θx significantly improved for the global parameters from 0.146 ± 0.115 mm ($\sum$ Local) to 0.086 ± 0.109 mm and 0.754 ± 0.496° ($\sum$ Local) to 0.483 ± 0.340°, respectively. Other metrics showed no significant differences.

Table 3.6: Quantitative outcomes of the optimized model evaluated on the test set, including the comparative analysis of the local and global parameters.

Final Model	**MAE** $(mean \pm sd)$				**FDR** $(mean \pm sd)$	**FD** $(median\ (min - max))$
	t_x (mm)	t_y (mm)	t_z (mm)	θ_x (°)	(%)	(mm)
Local	0.046 ± 0.028	0.011 ± 0.008	0.113 ± 0.066	0.072 ± 0.037	11.80 ± 10.31	5.024 ± 3.974
Global	0.904 ± 0.763	0.086 ± 0.109*	2.651 ± 1.540	0.483 ± 0.340*	11.63 ± 8.63	4.882 ± 2.873
∑ Local	0.844 ± 0.782	0.146 ± 0.115	2.630 ± 2.002	0.754 ± 0.496	11.80 ± 10.31	5.024 ± 3.974

* indicates a significant difference with the cumulative local parameters (∑ Local), with a p-value < 0.05 after Holm's correction.

3.2.3. Inference and visualization of final model

To illustrate the value of the optimized model, inference and visualization was applied on a sample from the test set. This sample was selected for its clear visibility of anatomical structures, which allowed for accurate manual segmentation. These results were visualized in three consecutive steps. Figure 3.3 presents a parameter-wise comparison of the predicted and ground truth transformation parameters, displaying both local and global labels. Additionally, the IMU orientation data are included for reference. The inference time for one sample was 0.17 seconds.

Subsequently, utilizing the predicted transformation parameters, the 2D US frames were transformed to reconstruct the probe's trajectory in 3D. Figure 3.4 shows the reconstructed trajectory with the US images positioned according to the predicted transformations. The ground truth trajectory is superimposed in green for comparison, highlighting the model's accuracy in spatial positioning.

To generate a continuous 3D reconstruction, voxel interpolation was applied to fill the gaps between frames using nearest neighbor interpolation. Figure 3.5 presents the resulting 3D volume, which took 67.8 seconds to reconstruct the US sequence of 115 frames. Manual segmentation of anatomical structures, including arteries and a colorectal tumor, was performed and incorporated into the 3D reconstruction using the predicted transformation parameters. In addition, a volume reconstruction of the 2D segmentations was performed using the average acquisition speed to determine t_z, and no transformations were performed for t_x, t_y, and θ_x. When this volume was placed in the 3D trajectory, it did not align with the anatomical features visible in the US reconstruction, while the volume reconstruction using the predicted transformation parameters demonstrated a high degree of similarity.

These visualizations translate the quantitative outcomes into practical demonstrations, underscoring the value of the accurate predictions of transformation parameters, for reconstructing the 3D trajectory and facilitating the assessment of anatomical structures, thereby highlighting the model's potential impact in clinical applications.

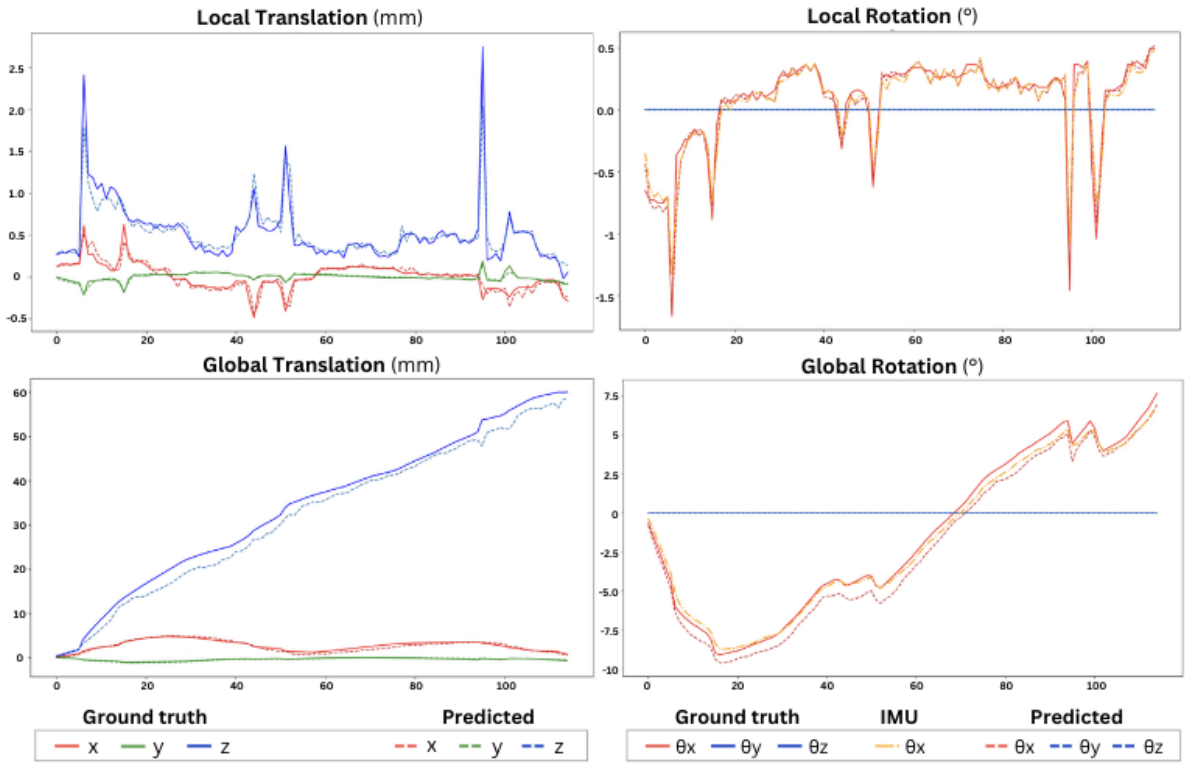


Figure 3.3: Inference sample of the test set, showing a parameter-wise comparison of the ground truth parameters and predicted parameters, with the IMU orientation for additional reference.

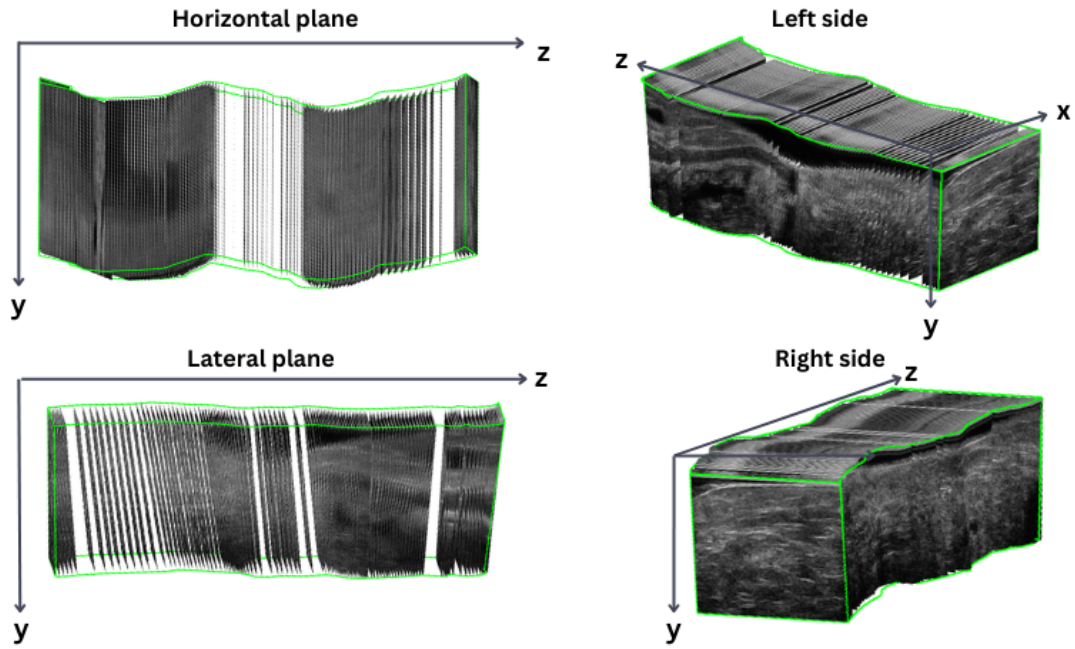


Figure 3.4: Spatial positioning of US frames in 3D space using the predicted transformation parameters, from four different viewpoints, with the ground-truth trajectory outlined in green.

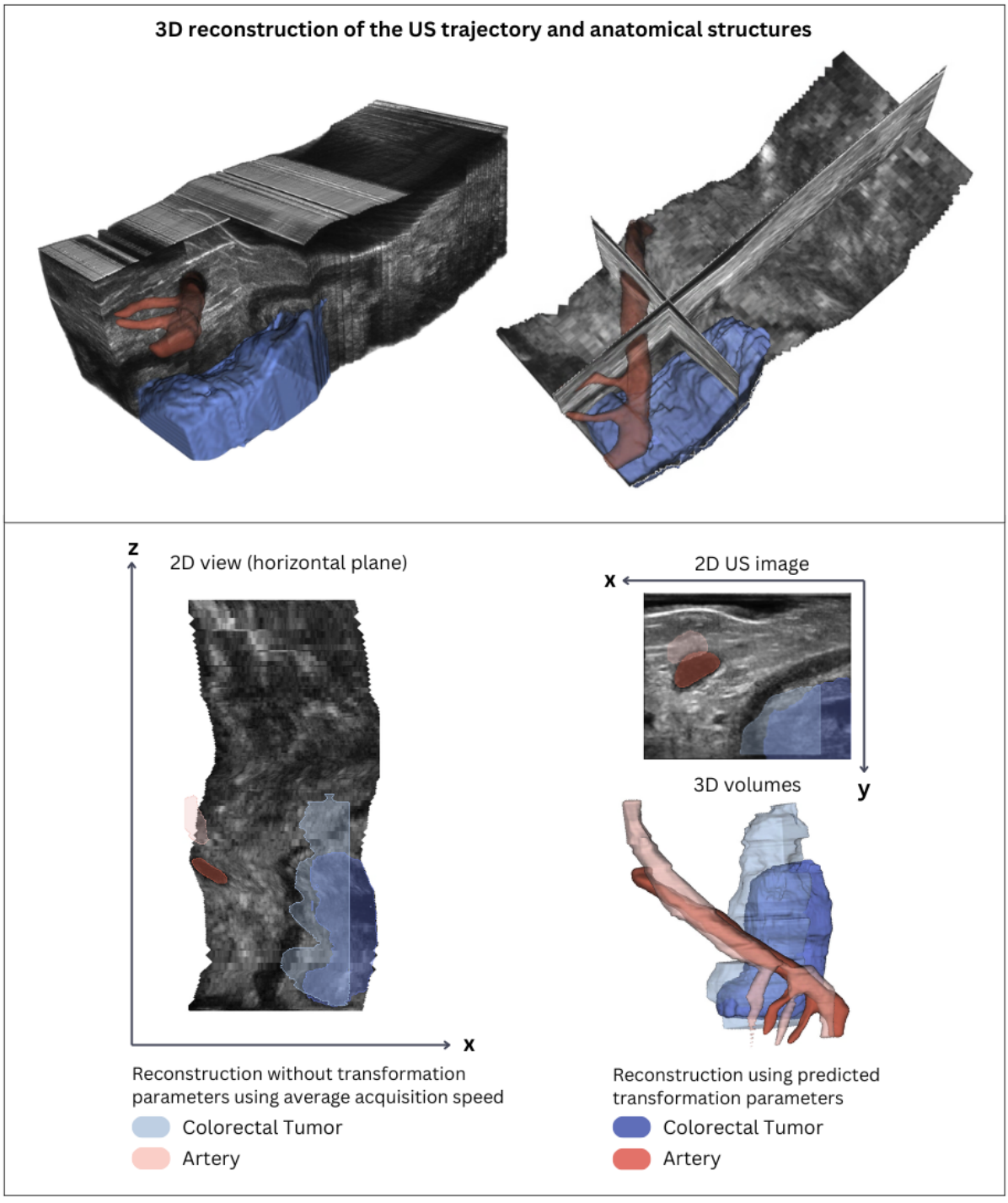


Figure 3.5: Reconstructed volumes of the US trajectory, and segmentations of a colorectal tumor and artery, using nearest neighbor interpolation. 3D volumes of the tumor and artery were reconstructed using the predicted transformation parameters, and without transformations for t_x, t_y, θ_x, and t_z determined based on the average speed.

4

Discussion

4.1. Interpretation of results

In this thesis, a deep learning-based 3D reconstruction model for POCUS imaging was developed, utilizing a newly acquired dataset and a novel network architecture combining a CNN with a transformer decoder. The dataset comprised 361 sweeps from ex-vivo specimens and a phantom model, collected using a customized measurement setup featuring a motorized scanner that facilitated ground truth positional tracking and movement of the US probe. Extensive data preparation and preprocessing resulted in an accurate aligned dataset of US images, optical flow and IMU data, all labeled with precise ground truth positions.

The baseline model demonstrated a reasonable accuracy in capturing local translations, with relatively accurate MAEs and a FDR of 10.46%. However, it exhibited substantial limitations in estimating global transformations, particularly the translation along the z-axis (t_z), with a large MAE of 6.606 ± 21.340 mm. This indicates that the model struggled to accurately capture complex scan trajectories involving out-of-plane movements, potentially leading to significant deviations during the sweep despite an acceptable FDR at the endpoint. These limitations were addressed by conducting four iterative ablation experiments successfully optimizing the model's performance. Most enhancements were attributed to Experiment 1, the integration of IMU orientation, and Experiment 2, employing sequence augmentation, with less improvement achieved in subsequent experiments.

Incorporation of IMU orientation, using the best-performing method, significantly improved the estimation of the orientation parameter θ_x, compared to utilizing US images and optical flow, from a global MAE of 1.342 ± 1.418° to 0.393 ± 0.295°. While improvements in local translation parameters were mixed, significant improvements were seen in overall global performance, indicating that the transformer benefits from IMU data by effectively learning temporal dependencies that enhance global trajectory predictions. Full potential of the IMU data was not completely leveraged, as only one rotational angle was incorporated in this setup. Therefore, with a future perspective of incorporating additional DOFs, or in the context of freehand scanning, integration of IMU data promises even more significance.

Applying data augmentation at the sequence level introduced greater variability in sweep trajectories, further reducing errors in MAE and FDR for both local and global parameters. This approach likely enhanced generalization by improving robustness to a wider range of scanning patterns. In the subsequent experiments, we selected the L2 loss function and maintained the dataset using subsampling with an interval of 5 frames, resulting in a final FDR of 9.00 ± 8.58%. The final model configuration

was selected based on the objective of reducing the trajectory error, with all performance metrics given equal weighting, according to the samples in this particular training set. It is important to note that the differences in outcomes between loss functions and datasets were minimal with high standard deviations. Consequently, an alternative selection in loss function or dataset could be considered an equally optimal choice.

Visualization of the optimized models over the ground truth trajectory yielded valuable insights for interpretation of the results, leading to noteworthy findings. In sequence 1 (Figure 3.2), the distance between frames was relatively consistent, and the final model achieved a FDR of 5.3%. In contrast, in sequence 2 (Figure 3.2), which exhibited greater variability in frame-to-frame distances, reflecting changes in speed, the model was less accurate (FDR of 9.8%). While the overall shape of the predicted sweep remained correct, the models tended to predict a shorter trajectory. In conclusion, in cases where the predictions were not perfect, the reconstruction still captured the general shape, including lateral deviations along the x-axis and accurate angles. However, due to the difficulties by the model in adapting to varying speed during the acquisition, reconstructions appeared slightly stretched or compressed.

When evaluated on the reserved test set, the final model exhibited a slight decrease in performance yielding a mean FDR of 11.63 ± 8.63%, which is expected when applying the model to unseen data. High standard deviations observed in both validation and test sets indicate variability in model performance across different cases. This can be attributed to the presence of outliers as well as inherent variability in the data, highlighting areas for future refinement. Nevertheless, the overall metrics and several highly accurate predictions on the unseen test set are considered to be promising.

Comparative analysis of the local and global parameters showed significant differences. Although the local parameters appeared slightly smoother in reconstruction, there is a significant improvement in t_y and θ_x using the global parameters, as evidenced by a reduction in MAE for t_y from 0.146 ± 0.115 mm to 0.086 ± 0.109 mm, and for θ_x from 0.754 ± 0.496 to 0.483 ± 0.340. This can likely be attributed to the transformer's ability to recognize global patterns specific to the scanning setup, such as the correlation between probe rotation and vertical movement. The capacity to learn temporal dependencies allows it to capture relationships that might not be apparent in local feature extraction. While global predictions are generated through the capability of global feature extraction by the Transformer, it is important to note that no conclusions can be drawn about the added value of the Transformer based on this comparison. A fair evaluation of the Transformer's contribution would require training with a different architecture that isolates the CNN and Transformer component and separates the losses.

4.2. Comparison to literature

In this study, we employed a novel data acquisition setup using a motorized scanner for ground truth positional tracking. Previous 3D US reconstruction studies predominantly utilized optical or EM tracking systems, often without reporting precise positioning accuracy. Some of them used optical systems intended for surgical navigation, which achieve resolutions around 0.2 mm [22]. However, many used EM systems with positioning resolutions of approximately 1.4 mm and orientation resolutions of 0.5° [21, 30, 31]. While widely adopted, these systems are prone to artefacts such as optical occlusion or drift and jitter due to electromagnetic interference from nearby metallic objects, resulting in limitations in accuracy [35]. By customizing the NEJE laser engraver for our purposes, we offer potential advantages in accuracy through precise control over probe movement, reducing reliance on external tracking systems susceptible to noise. This is particularly important given the critical need for reliable, low-noise data when dealing with minor frame-to-frame transformations. The motors in our setup have a travel resolution of 80 steps per millimeter, with NEJE specifying point positioning accuracy of 0.075 mm [36]. While this precision is promising for reliable positional tracking, the exact accuracy of our complete setup requires further validation, including direct comparisons with established optical and EM tracking technologies to substantiate its advantages.

In previous studies on 3D US reconstruction, deep learning models have predominantly utilized CNNs, including 2D CNNs like ResNet and EfficientNet [37]. Efforts to incorporate temporal dependencies have involved employing 3D CNNs such as ResNext to process temporal sequences [32, 38], integrating recurrent architectures like long short-term memory (LSTM) networks along with various consistency losses [21, 30, 31], or utilizing recurrent neural networks (RNNs) [37]. However, no consensus has emerged on the optimal architecture for modeling temporal information in US sequences. Recognizing limitations such as the sequential processing constraints and memory limitations of RNNs, and the limited temporal context provided by CNNs alone, we introduced a novel CNN-Transformer architecture for 3D trajectory reconstruction, effectively capturing global contextual information without the drawbacks of sequential processing inherent in RNNs. This could be particularly advantageous in longer US sequences, where drift accumulates, while the transformer may capture nuanced temporal changes more effectively. However, up until now this approach has remained unexplored in this context. The results of the current study demonstrate its feasibility for 3D US reconstruction.

It is essential to consider differences in datasets and trajectory complexities when comparing quantitative results with other studies, as these factors significantly impact performance metrics. Consequently, absolute MAE values are challenging to compare directly across studies. Only two studies reported parameter-wise MAEs, both demonstrating the highest errors in t_z compared to the equally high MAEs of t_x and t_y [22, 24]. This higher error in t_z aligns with our findings that estimating out-of-plane parameters is inherently more complex. In our case, the MAE for t_y is much lower due to minimal variation along this axis in our setup, but is expected to increase to the level of t_x, if we extend our setup further.

The FDR, being normalized by the length of the sweep, provides a more suitable metric for quantitative comparison across studies. As presented in Figure 4.1, state-of-the-art studies employing deep learning-based 3D trajectory reconstruction reported values with mean FDRs ranging from 9.64% [32] to 15.67% [30], and one study reported only the median FDR of 5.2% [22]. In our study, presented at the right side, we achieved a median FDR of 6.98% on the training set using five-fold cross-validation, and 9.00% on the unseen test set. Notably, the superior median FDR of 5.2% was achieved by Prevost et al. on a dataset of 600 sequences, twice the size of ours, collected from the forearms of 15 subjects, using two-fold cross-validation [22]. Additionally, the data from Luo et al.(2023b) consisted only of linear sweeps performed by a robotic arm, encompassing less variation in motions [21]. Therefore, considering our novel architecture and setup, our results demonstrate competitive performance with the state-of-the-art values reported in the literature.

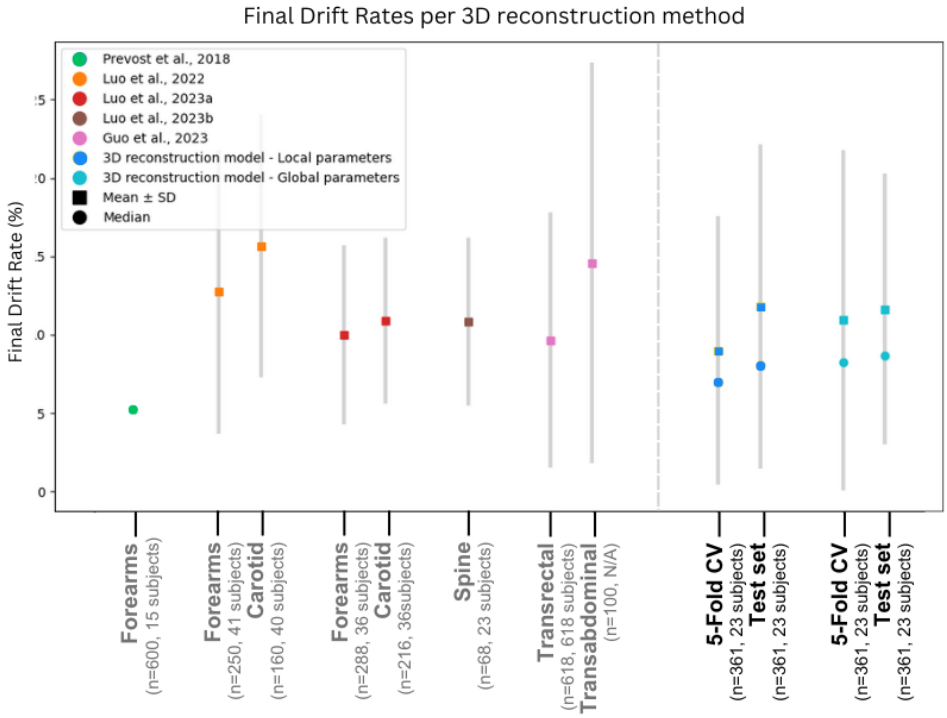


Figure 4.1: Reported mean and/or median Final Drift Rates (%) per reconstruction method across different datasets in literature (left), which are compared to the FDRs achieved in this study (right), with 'n' denoted as the number of sequences in the dataset.

4.3. Limitations and future recommendations

4.3.1. Data acquisition set-up and data collection

Despite the promising results achieved in this study, several limitations inherent to the data acquisition setup and data collection process are acknowledged. While the custom setup was instrumental in facilitating accurate positional tracking and data acquisition, providing precise data alignment and synchronization, it also imposed certain constraints on the developed reconstruction model, highlighting areas for future improvement.

Firstly, the application of the developed 3D reconstruction model is currently limited to the specific configurations of the data acquisition setup. The existing setup primarily facilitates translations along the x-axis (t_x) and z-axis (t_z), and rotations around the x-axis (θ_x), with minimal variations in the y-axis (t_y) resulting indirectly from rotational movements. Although this captures the primary movements employed during scanning of a resected specimen, the model was neither trained nor validated on transformation parameters such as roll and yaw (θ_y and θ_z) other than 0, or substantial variations in t_y. Additionally, the scanning trajectories, while including meandering paths and loops, were restricted to forward scanning without backward movements or repeated coverage of the same areas. This reduces the model's generalizability to more complex scanning patterns potentially encountered in freehand US examinations.

Moreover, practical limitations of the setup were observed in the analysis of local transformation predictions. Specifically, the 2D analysis revealed simultaneous peaks in local transformations across all axes (Figure 3.3). While optical and electromagnetic tracking systems also introduce noise and alignment challenges, we hypothesize that these peaks may be caused by the discrete movements of the stepper motors. The operation of the motor is not yet entirely smooth and seamless, occasionally manifesting as minor fluctuations or larger jumps in the local parameters. This often occurs simultaneously along all axes, as the G-code commands used to control the motors consist of three variables

(x, z, θ_x), consequently initiating motion along all axes at the same time. However, these peaks are minor and not discernible to the naked eye nor reflected in the global parameters.

To address these limitations, future research should focus on two main areas. First, enhancing the data acquisition set-up by introducing more DOF and incorporating a wider variety of scanning trajectories to provide a more comprehensive dataset. While the current dataset contains a substantial number of frames, the number of sequences (n=361) may be unsatisfactory for training a deep learning network to generalize across varied scanning patterns. As sequence level augmentation has shown favorable results by introducing more variability, this may also help reduce the amount of outliers observed [37]. Second, validation of the model on freehand scanning data in real-world clinical settings is essential to assess its applicability in clinical practice. It is important to determine how representative the trained model is when applied to data acquired outside of the controlled setup. In this regard, consideration should be given to the requirements of the intended clinical application; for instance, whether adherence to a specific scanning protocol is feasible or whether the model must accommodate the full variability of freehand scanning.

4.3.2. Development and optimization of 3D reconstruction model

The second category of limitations pertains to the development and optimization of the 3D reconstruction model, which were limited by time constraints. A CNN-Transformer architecture was selected based on its promising performance in handling temporal sequences. However, alternative architectures might be equally suitable or could offer improved performance.

While this architecture showed potential for our problem and yielded promising results, it also imposed certain limitations on the training implementation. Specifically, the Transformer component restricted the batch sizes that could be used during training, which might have constrained the local feature extraction by the CNN component. Alternatively, employing other architectures that are more computationally efficient could alleviate this issue. A careful trade-off analysis between the added value of the Transformer and the potential benefits of alternative approaches is warranted.

Moreover, while we experimented with common loss functions such as L1 and L2 losses, the complexity of the problem, incorporating four different types of predictions, suggests that other loss functions or strategies could be beneficial. Since errors can accumulate over the length of a sweep, the impact of errors may increase towards the end of a trajectory for global parameters. Therefore, it may be advantageous to consider percentage-based loss functions or those that account for cumulative error. Incorporating smoothness constraints into the loss function could also promote more realistic and physically plausible movements by penalizing abrupt changes in the predicted transformations.

Additionally, employing auxiliary losses or multi-task learning strategies could enhance model performance by providing additional guidance during training. Introducing intermediate supervision through loss functions at different layers of the network, for instance separating local losses from global losses, can help the model learn more meaningful feature representations that benefit the final prediction. Furthermore, predicting motion dynamics such as velocity, potentially by leveraging the currently excluded IMU accelerometer data, could provide valuable context for the transformations and improve the model's ability to capture temporal dependencies. Exploring these alternative loss functions and training strategies holds promise for further improving the accuracy and robustness of the 3D reconstruction model.

4.4. Clinical Perspective and Future Directions

Visualization of the reconstructed trajectories and volumes underscores the practical significance of accurate transformation estimation and the potential of the developed model. Without precise transformation parameters, reflected by the trajectory errors, the reconstruction and segmentation of anatomical volumes would be compromised. In Figure 3.5 it was shown that the predicted reconstruction enables correct segmentation of anatomical structures, crucial for clinical applications such as tumor localization and surgical planning.

While optimization of the 3D reconstruction model primarily aimed at reducing trajectory errors, it is important to recognize that minimizing trajectory errors is not an end goal in itself but rather a means to improve clinical outcomes. The level of accuracy required, measured in terms of MAE and FDR, largely depends on the specific clinical application. Encouragingly, in cases where the model demonstrated high performance, the current model may already be sufficient for certain clinical applications. However, to ensure reliable performance across all cases, it is crucial to first reduce the occurrence and impact of outliers.

Therefore, now that the technological feasibility of 3D US reconstruction with the current data acquisition setup has been demonstrated, aligning optimization objectives with clinically relevant outcome measures should be a priority in future research. This will enable the model to be optimized to meet specific clinical needs rather than solely focusing on trajectory reconstruction. Next to accuracy, this involves prioritizing factors such as computational efficiency, real-time processing capabilities, and the feasibility of implementing scanning protocols versus completely freehand scanning in clinical settings.

For instance, although we found that utilizing datasets with smaller subsampling intervals reduced performance in one metric in Experiment 4, there are substantial reasons to select such datasets when considering visualization purposes, as they provide more detailed information for voxel reconstruction. However, this comes at the cost of increased computational efforts. Similarly, while we observed that using the L2 loss function improved drift errors, it is important to consider whether minimizing drift is the most clinically relevant objective. An alternative approach would be to evaluate the tracking error of the ROI, rather than the frames within the entire trajectory, which may offer more clinically relevant insights. Furthermore, if the objective is to measure the volume as a crucial parameter for diagnosis, the Dice coefficient is paramount [39].

By tailoring the development of the model to specific clinical applications, such as quantitative volume measurements, tumor margin assessment, or integrating US reconstructions with advanced imaging modalities like MRI or CT, further improvements can be achieved, maximizing the model's impact on patient care. Close collaboration with clinicians is essential to define relevant outcome measures, explore alternative architectures and loss functions that better capture clinically significant aspects of the reconstruction, and balance computational efficiency with the level of detail required for clinical decision-making. In conclusion, while the current work represents a promising foundation for 3D trajectory reconstruction using deep learning, future efforts should prioritize clinical applicability by aligning model optimization with specific clinical needs and outcomes.

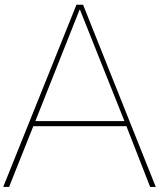
5

Conclusion

In this thesis, we addressed the challenge of reconstructing 3D US volumes without relying on external tracking devices or specialized 3D transducers, aiming to enhance applicability in clinical and point-of-care settings. A high-quality dataset was acquired using a POCUS probe with integrated IMU, mounted on a custom-designed motorized scanner. This setup facilitated precise positional tracking and movement control, allowing for the collection of 361 US sweeps from ex-vivo specimens and a phantom model.

This dataset facilitated the development of a CNN-Transformer architecture, leveraging 2D US images, optical flow, and IMU orientation data. Visual inspection and quantitative evaluations demonstrated that the model accurately captured the general shape and lateral deviations of the sweeps, reflected in low MAEs on the test set for translations along the x-axis (global MAE of 0.904 ± 0.763 mm), y-axis (global MAE of 0.086 ± 0.109 mm), and rotation around the x-axis (global MAE of 0.483 ± 0.340°). While the estimation in the scanning direction (t_z) showed a higher MAE of 2.651 ± 1.540 mm, reflecting challenges in adapting to varying speeds, this level of accuracy is still commendable given the complexity of out-of-plane motion estimation. The optimized model achieved a mean FDR of 10.94% on the training set using five-fold cross-validation and 11.63% on the unseen test set, with median FDRs of 6.98% and 8.10%, respectively, demonstrating competitive performance with state-of-the-art methods reported in the literature. Moreover, the predicted reconstructions enabled correct segmentation and visualization of anatomical structures in 3D, which is crucial for clinical applications such as tumor localization and surgical planning.

Consequently, this work demonstrates the feasibility and potential of combining advanced data acquisition techniques with a CNN-Transformer network for 3D US reconstruction. Although further validation and refinement of the methods is required, this work represents a significant contribution to improving the accuracy of 3D reconstruction, ultimately taking a step toward the adoption of trackerless 3D US in clinical and point-of-care environments.



Configuration NEJE motherboard during measurements

Table A.1: Parameter settings for the system configuration, in which '$' serves as prefix for the systemconfigurations and is used to configure or retrieve specific parameters in GRBL.

$ Nr	Parameter	Value
$0	Step pulse time, microseconds	10
$1	Step idle delay, milliseconds	255
$2	Step pulse invert, mask	0
$3	Step direction invert, mask	1
$4	Invert step enable pin, boolean	0
$5	Invert limit pins, boolean	0
$6	Invert probe pin, boolean	0
$10	Status report options, mask	3
$11	Junction deviation, millimeters	0.010
$12	Arc tolerance, millimeters	0.002
$13	Report in inches, boolean	0
$20	Soft limits enable, boolean	0
$21	Hard limits enable, boolean	0
$22	Homing cycle enable, boolean	1
$23	Homing direction invert, mask	1
$24	Homing locate feed rate, mm/min	250.000
$25	Homing search seek rate, mm/min	2500.000
$26	Homing switch debounce delay, milliseconds	250
$27	Homing switch pull-off distance, millimeters	1.000
$30	Maximum spindle speed, RPM	1000
$31	Minimum spindle speed, RPM	0
$32	Laser-mode enable, boolean	1
$40	Laser Focus Brightness	0.200
$41	Laser Temperature Auto-Reporting Interval	10
$42	Auto-Sleep Time	10
$43	Tilt Detection Sensitivity	0
$100	X-axis travel resolution, step/mm	80.000
$101	Y-axis travel resolution, step/mm	80.000
$102	Z-axis travel resolution, step/mm	800.000
$103	A-axis travel resolution	8.889
$110	X-axis maximum rate, mm/min	15000.000
$111	Y-axis maximum rate, mm/min	15000.000
$112	Z-axis maximum rate, mm/min	1200.000
$113	A-axis travel resolution	21600.000
$120	X-axis acceleration, mm/sec^2	250.000
$121	Y-axis acceleration, mm/sec^2	250.000
$122	Z-axis acceleration, mm/sec^2	20.000
$123	A-axis travel resolution	250.000
$130	X-axis maximum travel, millimeters	170.000
$131	Y-axis maximum travel, millimeters	170.000
$132	Z-axis maximum travel, millimeters	45.000
$133	A-axis travel resolution	360.000

B

Optical flow parameters

Table B.1: Parameters used for denoising of ultrasound images and optical flow computation

Denoising Step	**Parameter**	**Value**
Gaussian Blur	Gaussian Kernel Size	(9,9)
	Sigma Value	0 (auto)
Bilateral Filter	Sigma Color	75
	Sigma Space	75
Sharpening	Sharpening Kernel	$\begin{bmatrix} 0 & -1 & 0 \\ -1 & 5 & -1 \\ 0 & -1 & 0 \end{bmatrix}$
Optical Flow Parameters	Pyramid Scale Factor	0.5
	Pyramid Levels	3
	Window Size	31
	Iteration Count	20
	Polynomial Neighborhood Size	9
	Polynomial Sigma	2

C

Processing IMU accelerometer data

The accelerometer data of the IMU have the potential to provide information about the translational movements of the probe during scanning by integrating over time to estimate velocity and displacement. However, a significant challenge with accelerometer data is the high level of noise, which often renders it unsuitable for accurate translation estimation. Therefore, effective preprocessing is essential to mitigate noise and address issues such as coordinate system alignment, ensuring the data is more reliable for subsequent analysis.

The raw data, which was extracted during data acquisition, is subjected to three processing steps. These include the normalisation of the data to account for its orientation, the implementation of denoising techniques and, finally, the integration of the data twice by time, which results in translation. The steps are described in detail below, and an illustrative example of a data sample that has undergone these steps is provided in Figure C.2.

1. Normalization for coordinate system

The accelerometer measures acceleration along three axes in its local device coordinate system, which therefore changes with the probe's orientation. To obtain acceleration readings in the global reference frame, it was necessary to account for the probe's orientation during scanning, in particular the varying tilting motion introduced by the motorized scanner. The orientation of the probe is illustrated in Figure C.1. Since the scanner only introduced rotation around the X-axis (pitch), applying a rotation matrix to the accelerometer data at each time point for rotation around the X-axis was sufficient. Therefore, a rotation matrix corresponding to the pitch angle (θx) was constructed to transform the accelerometer data into the global reference frame:

$$R_x(\theta_x) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta_x & -\sin\theta_x \\ 0 & \sin\theta_x & \cos\theta_x \end{bmatrix}$$

After correction for its variable rotation around the x-axis, the accelerometer data were normalized to account for any initial biases or offsets in orientation of roll, pitch and yaw. This was achieved by subtracting the mean of the accelerometer readings acquired during its initial stationary position prior to the movement of each sweep. This step removed any constant acceleration components (e.g., due to gravity or sensor bias) and set the baseline acceleration to zero.

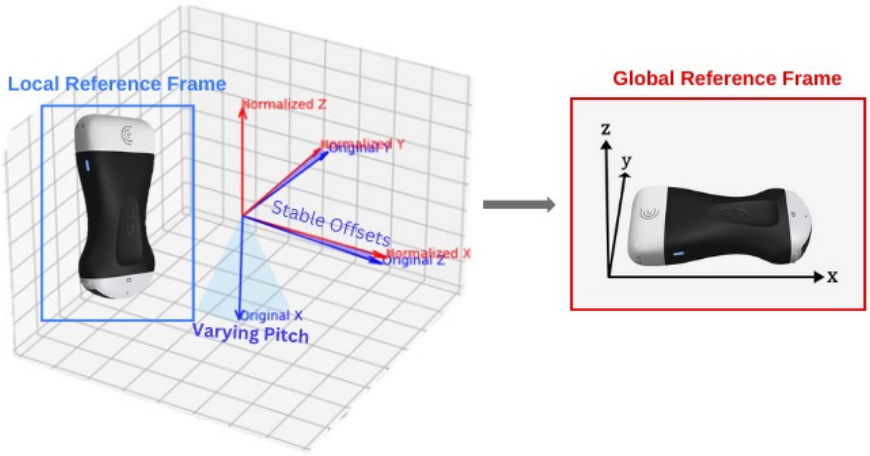


Figure C.1: Example of normalization for intial position by varying pitch during using the transformation matrix, and stable roll and yaw using initial offset visualized in green

2. Denoising Strategies

The accelerometer readings were susceptible to high levels of noise due to factors such as electrical interference, mechanical vibrations from the stepper motors, and the inherently low acceleration levels associated with the scanner's slow movement speeds (1.8–2.6 mm/s). These low acceleration signals could be difficult to distinguish from the sensor's noise floor, the level of background noise inherent in the accelerometer's output, even when no acceleration is applied. To improve the data quality, several denoising techniques were considered. Low-pass filtering aimed to remove high-frequency noise while preserving the low-frequency signals corresponding to actual movements. Band-pass filtering was explored to isolate the movement-related frequencies by removing both low-frequency drift and high-frequency noise. Lastly, Kalman filtering, an adaptive method accounting for measurement noise and system dynamics, was also attempted.

3. Integration over time

Following denoising, the accelerometer data was integrated twice over time to estimate velocity and displacement. To compare the derived displacement, the estimated ground truth position of the IMU sensor was calculated based on the motor positions using the spatial offsets of the IMU components with respect to the center of the imaging array, as described in Section 2.1.3. While these ground-truth IMU positions were not used as labels for model training, they served as a reference for evaluating and comparing denoising strategies applied to the accelerometer signals.

Conclusion

Despite several preprocessing steps, significant challenges were encountered in obtaining reliable displacement estimates from the accelerometer data. Integration of residual noise led to substantial drift over time in the displacement estimates, exacerbated by the low signal-to-noise ratio due to the slow scanning speeds. The very small acceleration signals were often indistinguishable from the sensor's inherent noise, making it difficult to extract meaningful motion information. Additionally, movement was initiated by stepper motors generating mechanical vibrations, which may have introduced high-frequency noise. This complicated the filtering process and further degraded signal quality. Given these challenges, it was determined that the accelerometer data were not suitable for providing accurate translational information in this context. Consequently, the accelerometer data were not included as input features for the deep learning model.

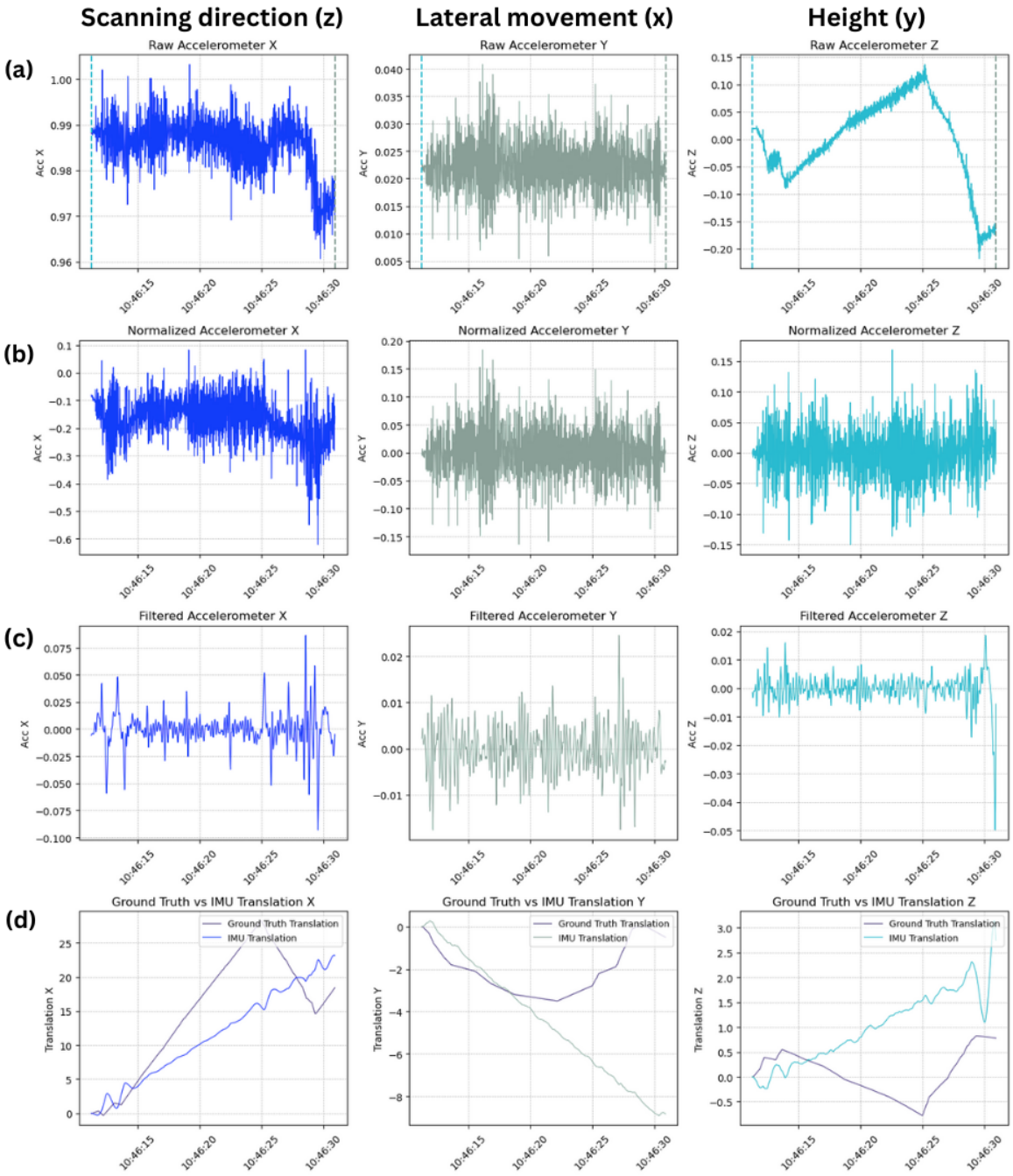


Figure C.2: Visualization of accelerometer data x, y and z for a US sequence over time. (a) Raw accelerometerdata, cropped to the ROI (movement of the sweep) (b) Normalized accelerometer data. (c) Filtered accelerometer data using bandpass filter. (d) Comparison of IMU translation (double integration of accelerometer data) with the ground-truth position of the IMU.

D

Ablation experiments

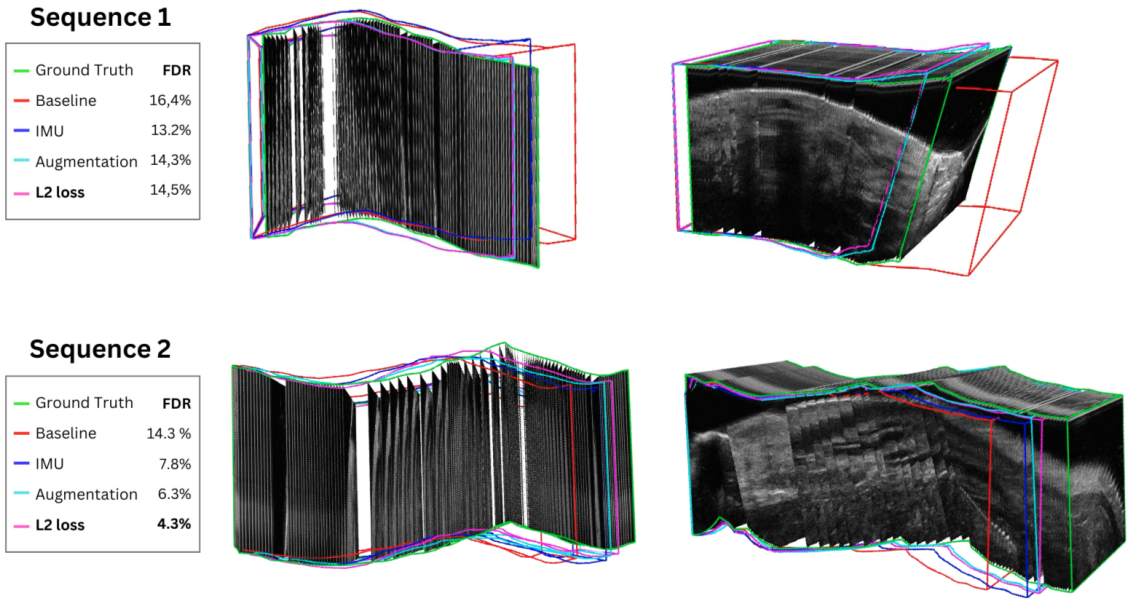


Figure D.1: Visualization of the optimized models against the baseline model and ground truth trajectory, based on the predictions of the local parameters.

Bibliography

[1] A. Fenster and D. B. Downey. “Three-Dimensional Ultrasound Imaging”. In: *Annual Review of Biomedical Engineering* 2.2000 (2000), pp. 457–475.

[2] M. H. Mozaffari and W.-S. Lee. “Freehand 3-D Ultrasound Imaging: A Systematic Review”. In: *Ultrasound in Medicine & Biology* 43.10 (2017), pp. 2099–2124.

[3] G. Unsgård, O. Solheim, F. Lindseth, and T. Selbekk. “Intra-operative Imaging with 3D Ultrasound in Neurosurgery”. In: *Intraoperative Imaging*. Ed. by M. N. Pamir, V. Seifert, and T. Kiris. Vienna: Springer Vienna, 2011, pp. 181–186.

[4] F. Makouei, T. D. Frehr, T. K. Agander, et al. “Feasibility of a Novel 3D Ultrasound Imaging Technique for Intraoperative Margin Assessment during Tongue Cancer Surgery”. In: *Current Oncology* 31.8 (2024), pp. 4414–4431.

[5] S. Dekalo, Z. Savin, E. Schreter, et al. “Novel ultrasound-based volume estimation of prostatic benign enlargement to improve decision-making on surgical approach”. In: *Therapeutic Advances in Urology* 13 (2021). PMID: 33633800, p. 1756287221993301.

[6] A. Hashim, M. J. Tahir, I. Ullah, et al. “The utility of point of care ultrasonography (POCUS)”. In: *Annals of Medicine and Surgery* 71 (Nov. 2021).

[7] S. Smith, H. Pavy, and O. von Ramm. “High-speed ultrasound volumetric imaging system. I. Transducer design and beam steering”. In: *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control* 38.2 (1991), pp. 100–108.

[8] A. Lang, V. Parthasarathy, and A. K. Jain. “Calibration of EM Sensors for Spatial Tracking of 3 D Ultrasound Probes”. In: *Data Acquisition Applications*. IntechOpen, 2012. Chap. 0.

[9] Q. Huang and Z. Zeng. “A Review on Real-Time 3D Ultrasound Imaging Technology”. In: *BioMed Research International* 2017 (2017), p. 6027029.

[10] *Clarius Website*. `http://www.clarius.me`. Web Page. 2024.

[11] J. F. Chen, J. B. Fowlkes, P. L. Carson, and J. M. Rubin. “Determination of scan-plane motion using speckle decorrelation: Theoretical considerations and initial test”. In: *International Journal of Imaging Systems and Technology* 8.1 (1997), pp. 38–44.

[12] R. F. Chang, W. J. Wu, D. R. Chen, et al. “3-D US frame positioning using speckle decorrelation and image registration”. In: *Ultrasound in Medicine & Biology* 29.6 (2003), pp. 801–812.

[13] A. H. Gee, R. James Housden, P. Hassenpflug, et al. “Sensorless freehand 3D ultrasound in real tissue: Speckle decorrelation without fully developed speckle”. In: *Medical Image Analysis* 10.2 (2006), pp. 137–149.

[14] T. Liang, L. S. Yung, and W. Yu. “On Feature Motion Decorrelation in Ultrasound Speckle Tracking”. In: *IEEE Transactions on Medical Imaging* 32 (2013), pp. 435–448.

[15] N. Afsham, A. Rasoulian, M. Najafi, et al. “Nonlocal means filter-based speckle tracking”. In: *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control* 62.8 (2015), pp. 1501–1515.

[16] Y. Lecun, Y. Bengio, and G. Hinton. “Deep learning”. In: *Nature* 521.7553 (2015), pp. 436–444.

[17] V. Ghasemzade and F. Jamshidi. “Applications of Inertial Navigation Systems in Medical Engineering”. In: *Journal of Physics and Biomedical Engineering* 8 (Jan. 2016).

[18] C. A. Adriaans, M. Wijkhuizen, L. M. van Karnenbeek, et al. “Trackerless 3D Freehand Ultrasound Reconstruction: A Review”. In: *Applied Sciences* 14.17 (2024).

[19] A. Vaswani, N. Shazeer, N. Parmar, et al. “Attention is all you need”. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS’17. Long Beach, California, USA: Curran Associates Inc., 2017, pp. 6000–6010.

[20] Clarius. *User Manual*. Accessed: 2024-11-06.

[21] M. Luo, X. Yang, H. Wang, et al. “RecON: Online learning for sensorless freehand 3D ultrasound reconstruction”. In: *Medical Image Analysis* 87 (2023).

[22] R. Prevost, M. Salehi, S. Jagoda, et al. “3D freehand ultrasound without external tracking using deep learning”. In: *Medical Image Analysis* 48 (2018), pp. 187–202.

[23] S. Balakrishnan, R. Patel, A. Illanes, and M. Friebe. “Novel Similarity Metric for Image-Based Out-Of-Plane Motion Estimation in 3D Ultrasound”. In: *Proceedings of the 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS)*. 2019, pp. 5739–5742.

[24] K. Miura, K. Ito, T. Aoki, et al. “Localizing 2D Ultrasound Probe from Ultrasound Image Sequences Using Deep Learning for Volume Reconstruction”. In: *Medical Ultrasound, and Preterm, Perinatal and Paediatric Image Analysis. ASMUS PIPPI 2020*. 2020, pp. 97–105.

[25] W. Wein, M. Lupetti, O. Zettinig, et al. “Three-Dimensional Thyroid Assessment from Untracked 2D Ultrasound Clips”. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*. 2020, pp. 514–523.

[26] B. K. Horn and B. G. Schunck. “Determining optical flow”. In: *Artificial Intelligence* 17.1 (1981), pp. 185–203.

[27] G. Farnebäck. “Two-frame motion estimation based on polynomial expansion”. In: *Image Analysis: Proceedings of the 13th Scandinavian Conference, SCIA 2003, Halmstad, Sweden, June 29 – July 2, 2003*. Springer Berlin Heidelberg, 2003, pp. 363–370.

[28] W. Baird. “An introduction to inertial navigation”. In: *American Journal of Physics* 77 (Sept. 2009), pp. 844–847.

[29] A. Howard, M. Sandler, G. Chu, et al. “Searching for MobileNetV3”. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2019, pp. 1314–1324.

[30] M. Luo, X. Yang, H. Wang, et al. “Deep Motion Network for Freehand 3D Ultrasound Reconstruction”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2022, pp. 290–299.

[31] M. Luo, X. Yang, Z. Yan, et al. “Multi-IMU with online self-consistency for freehand 3D ultrasound reconstruction”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2023, pp. 342–351.

[32] H. Guo, H. Chao, S. Xu, et al. “Ultrasound Volume Reconstruction From Freehand Scans Without Tracking”. In: *IEEE Transactions on Biomedical Engineering* 70.3 (2023), pp. 970–979.

[33] F. Wilcoxon. “Individual Comparisons by Ranking Methods”. In: *Biometrics Bulletin* 1.6 (1945), pp. 80–83.

[34] S. Holm. “A Simple Sequentially Rejective Multiple Test Procedure”. In: *Scandinavian Journal of Statistics* 6 (1979), pp. 65–70.

[35] A. M. Franz, T. Haidegger, W. Birkfellner, et al. “Electromagnetic Tracking in Medicine—A Review of Technology, Validation, and Applications”. In: *IEEE Transactions on Medical Imaging* 33.8 (2014), pp. 1702–1725.

[36] NEJE Laser Engraving Wiki. *NEJE 3 Documentation*. Accessed: 2024-11-05. 2024.

[37] Q. Li, Z. Shen, Q. Li, et al. “Long-Term Dependency for 3D Reconstruction of Freehand Ultrasound Without External Tracker”. In: *IEEE Transactions on Biomedical Engineering* 71.3 (2024), pp. 1033–1042.

[38] H. Guo, S. Xu, B. Wood, and P. Yan. “Sensorless Freehand 3D Ultrasound Reconstruction via Deep Contextual Learning”. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*. 2020, pp. 463–472.

[39] A. A. Taha and A. Hanbury. “Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool”. In: *BMC Medical Imaging* 15 (Aug. 2015).