Tomographic Reconstruction of Freehand Non-Tracked Ultrasound Data

Using Deep-Learning and Transducer-Specific Attention Maps

Utrecht University

Tiziano Natali
Information and Computing Science
Netherlands Cancer Institute - Utrecht University

Game and Media Technology
Msc. Thesis
June, 2021

Contents

Chapter 1

Introduction

Every year, more then 900 thousands people are diagnosed primary liver cancer or colorectal liver metastases ([38]). Currently different types of treatments for liver cancer are available and they consist in liver resection, minimally invasive interventional procedures (RadioFrequency Ablation or radio-embolization), radiotherapy, systemic therapy ([33]) or a combination of the aforementioned techniques. Although each of these methods have its pros and cons, when feasible liver surgery (in the form of either ablation or resection) is preferred, due to the best long-term patient prognosis ([41]).

In the case of liver surgery, planning and execution of surgical resection is especially challenging, since the organ is characterized by a complex underlying hepatovascular and biliary anatomy, which varies from patient per patient. As a result, incomplete resections of the tumor or to intra/post-operative complications can often happen. Since the liver is so deformable, its shape during surgery can be drastically different from how it looked in a preoperative scan (usually a CT or MRI scan). Therefore, intraoperative localization of the target lesion should be facilitated by the addition of high quality tomographic images into the surgical environment (e.g., for 3D positioning of the lesions). Giving the physician a 3D visualization of the available information of the target lesion can facilitate the surgery execution.

During the course of the last decades, research has focused primarily on developing ways to facilitate medical diagnosis. Many tools have been developed to facilitate preoperative patient-specific resection planning (e.g., 3D modelling of patient anatomy, automatic segmentations). As a result, the whole surgical team can carefully select patient suitable to surgical resection, and plan the optimal resection plan. However, when it comes to actual intraoperative execution of this plan, there has not been much development,therefore surgeons will need to primarily rely on their recollection of the preoperative plan and basic palpation of the organ, to steer the resection in the right way.

There are several ways to incorporate tomographic images in the surgical environment for better localization of the target lesion. The most widely available and used tool to perform intraoperative imaging is the ultrasound (US) scanner. Even though these systems are usually cheap, and present in the vast majority of the surgery rooms, it can be hard for a surgeon to orient him/herself in

the 3D anatomy of the scanned organ using a single 2D cross-section (e.g., visualized with US). In contrast to radiologists, surgeons are not trained to orient themselves in 3D from a 2D US image. Therefore, this might elongate the surgery time which is usually undesired.

A second approach is to perform an intraoperative operative tomographic imaging (e.g., CT, CBCT or MRI) of the liver. This technique generates high quality images and there are many ways to generate high fidelity 3D reconstructions from the scan. Although the quality the reconstructions from CT/CBCT/MRI scans is great, the procedure requires expensive specific instruments that are not available in the majority of the surgery rooms (e.g., imaging is restricted to hybrid OR only). Also, it adds time to the surgery procedure to perform the scan on the patient, and imaging is only possible at given time-point during the resection (e.g., no continuous imaging).

Another approach to incorporate information about the target lesion and generate a 3D visualization can be achieved using 2D US images. There are several options that can be used to generate 3D visualizations from US images. An immediate one is the implementation of a 4D US transducer. These transducers can generate live, high-quality 3D reconstructions of the scanned area. However, these transducers are expensive and large, which makes them unsuited for liver surgery intraoperative scans. When performing a sweep (the procedure of passing once the transducer over the region of interest) on the liver during surgery, it can be required to place the transducer inside of the patient (e.g., direct contact with the organ). Therefore, the transducer has to necessarily be as small as possible, to be able to reach distant and restricted cavities. Consequently, 3D US transducers are not suited for intraoperative imaging in liver surgery.

Another approach to generate 3D visualization from US data is to use tracked 2D US transducers. When performing a sweep, the transducer scans the Region Of Interest (ROI) and acquires a series of 2D images. From the stack of 2D images generated from the US sweep, it is possible to generate a 3D volume. However, the motion of the US transducer during a freehand sweep is non-linear. For example, the transducer can be moved at different speed or can be used with different pressure on the liver, which causes the organ to deform. All of these variables cause an uneven sampling of the US acquisition. Therefore, generating high fidelity 3D volumes from the 2D US images is not a straightforward task. Mechanical-driven robotic US arms can overcome the linearity of the acquisition. However, these systems are usually large thus not suited for intraoperative liver surgery.

The implementation of systems whose aim is to track the position and rotation of the transducer during the course of the US sweep can help with the 3D visualization (or reconstruction) of 2D US images. It is possible in a second time to generate a 3D reconstruction of the scanned ROI using the tracked positions and rotations.

There are two kinds of tracking systems: optical and electromagnetic. The first one, which is also the most accurate, tracks the trajectory of a reflective optical marker on the transducer by the means of a stereo camera. This kind of tracking system is not suited for intraoperative liver surgery, since optical tracking requires constant line-of-sign agreement between the marker and the camera, but this is not always possible during liver surgery due to occlusion (by the physicians or the patient). Another challenge is relative bulkiness of optical markers, what, similar to 4D transducers, restrict organ accessibility with optically tracked transducers.

In contrast, electromagnetic (EM) tracking is based on passive tracking micro-EM sensors position with artificially generated EM-field. This artificial filed is generated by external filed generator, that are place under or within close proximity of the patient, and are able to cover the whole resection field. EM sensor itself represent a small (1x8mm) wired beacon, that could be placed on the top of the US transducer and could be tracked in 3D (6 degrees of freedom tracking). The receiver perturbates the EM field and its position and rotation are tracked. These systems are the most suited for US reconstruction during liver surgery, since they do not have line-of-sight limitation. Still, adopting this kind of systems is not ideal in a real surgery room scenario. These systems are usually expensive and technical support is required in the surgery room to make them work. Moreover, tracking accuracy of EM-sensor could be easily influenced by interference with metallic objects or active EM field (e.g., ablation needle). In general, accuracy of EM-tracked US has been proved to be good with rigid body parts (e.g., limbs, neck) ([8]) but their use on the reconstruction of deformable organs is not documented. When performing an US sweep, the transducer has to keep continuous contact with the scanned area. If the transducer is not in contact with the scanned area, no signal will be recorded (e.g., air) and the image will be blank. Therefore, it is required to put pressure on the organ and it can occur that the organ gets deformed. If the pressure on the organ is not constant, the look of the anatomy will change, hence the reconstruction might suffer from generated artifacts in the final visualization.

Consequently, the research in the field of 3D US volume reconstruction has focused on developing a sensorless freehand scanning system, trying to overcome all of the difficulties that it brings. These systems usually are a two-step process: tomographic reconstruction from US images ([42][4][32]) and volume computation ([46][22][18]). The premises listed above lead to the goal of this study, which will be explained in the next section.

1.1 Motivation and Research Question

This study aims at understanding how to avoid the use of expensive tracking systems to generate a 3D volume from a standard 2D ultrasound sweep. The final output of the study is a new tomographic reconstruction method from freehand 2D US data without tracking, based on a deep-learning model.

The main aspects that will have to be taken into an account during development of the the method are: (i) the method will be used for medical purpose, thus the generated volumes have to be as close as possible to reality; (ii) the target maximum error will be limited at 5 millimeters with respect to the ground truth (e.g., clinically acceptable range).

Given these aims and requirements, the study tries to answer the following question.

Research Question: is freehand tomographic reconstruction from untracked 2D US data in intraoperative ultrasound for liver surgery feasible?

Since this is a broad research question, this study will try to answer the three following sub-questions:

1. Is it feasible to adapt the tomographic reconstruction method for US imaging of extremities proposed by Prevost *et. al.* [32] for imaging of mobile and deformable anatomies (liver)?

2. Will addition of transducer-specific geometry attention map be able to improve geometric accuracy of the reconstruction method, developed to answer sub-Question 1?

3. Will multi-task learning have beneficial influence on performance of tomographic reconstruction method developed in sub-Question 1?

To answer sub-Question 1, the same metrics adopted in [32] will be used for a direct comparison:

- **Average parameter-wise absolute error**, the estimation error between ground truth and generated displacement vector averaged over all images in the volume.
- **Drift**, the distance between the last ground truth and the last generated translation vector.
- **Length error**, the difference in length of the ground truth and estimated volume.

Thereafter, the results found from sub-Question 1 will be used in the comparison between the base model and its. The results of these comparisons will give an answer to sub-Questions 2 and 3.

The remainder of this report is organized as follows. In Chapter 2, we will be described the previous works and the literature relevant to the purpose of this study. Thereafter, an overview of the methodologies and the available data will be described in Chapter 3. Chapter 4 will present the results and the experiment used to acquire them, discuss the limitations of the models and provide conclusive thoughts.

Chapter 2

Related Work

In the following sections will be presented literature and previous works relevant for this study. First, the different tracking systems for ultrasound scans will be reviewed. Thereafter, the most relevant literature in the topic of ultrasound registration will be presented followed by an overview of the previous works on tomographic reconstruction from 2D US images.

2.1 Scanning Systems

During a standard ultrasound sweep, the physician analyzes the images and has to mentally reconstruct conventional 2D US images into a 3D model of the ROI (e.g., a body part or an organ). This is a challenging step, that often requires repetitive sweeps of the ROI to get a proper evaluation of the organ. During a surgery, additional surgical delays are strongly undesired, therefore several setups were developed to simplify US-based assessment of complex organs (e.g., liver).

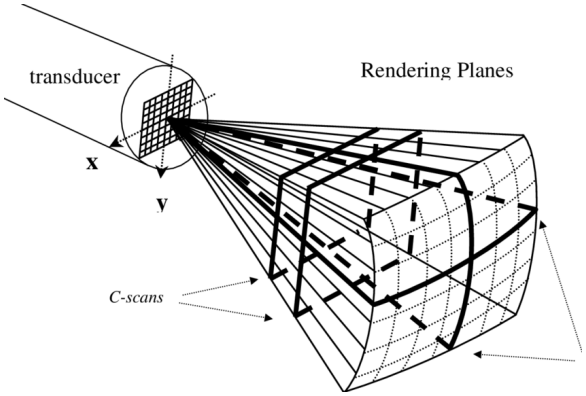


Figure 2.1: Schematic of the principle of volumetric imaging with 2D array transducer. Source [17]

In order to simplify intraoperative assessment of the liver using US, two methods could be used. First, 4D phase array US transducers could be used to generate live 3D images. However, their large size makes them unsuited for intraoperative operations. Second, a stream of conventional 2D US images could be reconstructed into a 3D volume. These approaches can be divided in the following major categories, depending on the method of volumetric image formation [26]: freehand scanning systems, tracked 2D scanning systems, volumetric scanning systems, mechanical scanning systems.

2.1.1 Freehand - "Video" Recordings of Sequential 2D US

In freehand US scanning systems, the physician directly operate the standard transducer, which generates a 2D ultrasound image of the scanned region. A standard US transducer has a 1D array of US emitters on the contact surface that can be placed in different ways. There are several kinds of US transducers designed for different tasks, which also emits uniquely-shaped US beams. Figure 2.2 shows the shapes of three of the most widely used US transdurcers.

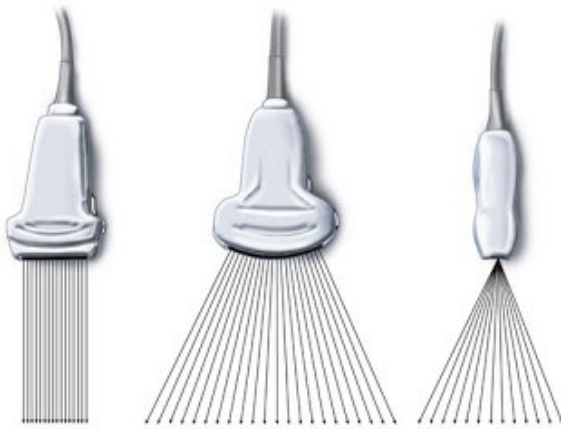
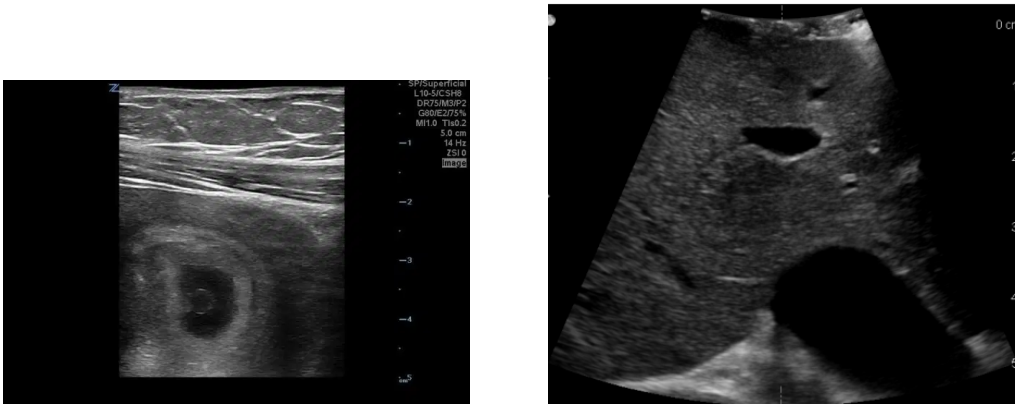


Figure 2.2: Shape of the US beam of the most widely used US transducers.

The most relevant aspect behind the different beams is the fact that, while the US image generated by a linear transducer will have relevant information in a square area, images originated by a convex transducer will have relevant information in a conical shape(see Fig. 2.3). Therefore, it is much easier to extract the core information from linear US images (e.g., simple cropping) then from convex US images.

Figure 2.4 schematically illustrates the basic principle of US image information. Piezoelectric elements positioned near the contact surface (the head) of the transducer emits ultrasound waves. Shape and distribution of these elements results in hourglass-like shape of ultrasound beam, as illustrated in the Figure 2.3. This means that the ultrasound beam has a focal zone where the waves scans a smaller area. When the image is reconstructed from the US waves, each pixel in the US image will have an intensity given by an average of the scanned area. Since in areas further from the focal point is averaged over wider region, this will influence image values that are produced in this area. On the other hand, close to the focal zone, the averaging will be done on an approximately uniform area (equal thickness), hence pixel values will be closer and more true to the ground truth.



(a) US image generated from a linear ultrasound transducer

(b) US image generated from a convex ultrasound transducer

Figure 2.3: US images generated from different US transducers. In contrast with the convex US image, relevant information can be cropped from the linear US image with no waste of information.

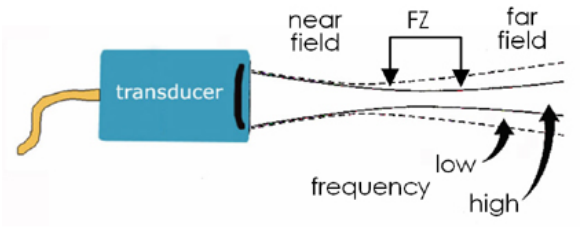


Figure 2.4: Focal zone of the beam emitted by an ultrasound transducer.

Another important characteristic of US transducers is their focusing method: fixed focus transducers, auto-focus transducers and manual focus transducers. As its name describes, fixed focus transducers are those, which always focus their US beam at a certain fixed distance. Therefore, even if the depth of the scanner is changed, the focal area will be fixed to a certain range. On the other hand, manual-focus transducers allows users to change the focus to the desired depth (e.g., varies acoustic lens of piezoelectric element). At last, auto-focus transducers does not give the possibility of focusing on a specific area. They instead auto focus to a fixed point in the 2D image, as illustrated on the screen, independent of the actual depth of the images. In this case, actual image resolution of the image illustrated on the screen of an US scanner stay always the same, while actual acoustic focusing settings of the transducer automatically adjust, based on user presets.

2.1.2 Tracked 2D

Most of the approaches can be assigned to one of the following categories: electromagnetically tracked systems (see Fig. 2.5b) and optically tracked systems (see Fig. 2.5a).

Electromagnetically tracked systems require specific hardware to operate. Usually they need

three main components to work: a field generator, a receiving sensor and a computer. The field generator is linked to the computer and emits an electromagnetic field. The sensor, which is attached to the transducer, generates a perturbation of the field. The generator is able to read the position and rotation of the sensor based on this perturbation. The computer stores and processes the acquired data through dedicated software. The system for navigated laparoscopic surgery proposed by Lango *et al.* [36] is based on this principle. A more recent work based on the setup proposed by Lango *et al.* is the one presented by Thompson [40] *et al.*. In their system they use an Aurora tracking system built by NDI Systems and CustusX [1] to read and store the data. In their work, the tracking stores the position and orientation of the transducer as offset and affine transformation matrix in relation to a reference coordinate system. Although these systems seems to solve the problem of tomographic reconstruction from 2D US data, their implementation is in fact far from ideal. These systems are expensive, they require extra hardware in the surgery, specific knowledge to be correctly operate from the surgeon and the support of a technician. Because of the aforementioned reasons, it is best to avoid using these systems for liver surgery.

(a) Example of optical tracking system for freehand US.

(b) Example of EM tracking system for freehand US.

Figure 2.5: Tracking systems for 2D freehand ultrasound produced by NDigital Inc..

In optical systems the transducer's position is tracked by means of a stereo camera. An optical reflective marker is positioned on the transducer and its position is tracked by cameras. Although this approach has great portability, it implies that the marker must be visible from the cameras at all time. During surgeries, it is not possible to ensure that this will happen. Also, the marker is usually large, which is inconvenient during US sweeps. Being the marker attached to the US transducer, its dimensions makes it harder (sometimes impossible) to reach far and narrow cavities in the abdomen of the patient. Two examples of optical tracking systems are the Polaris and the Optotrak Certus, both from NDI (Northern Digital Inc., Waterloo, Canada). To avoid the marker tracking problems, some systems tried to add a inertial sensor that could track the position of the transducer when the

cameras are occluded. In their work [25] Mohamed et al implemented an optical-inertial tracking system using the PlayStation Move as inertial sensors. Although this tries to solve the occlusion related issues, the size of the markers and of the additional inertial sensors still remains a problem.

2.1.3 Volumetric

A volumetric scanning systems refers to a single volumetric US transducer containing a large 2D phase array of piezo elements. Considering large size of the 2D matrix array, it is possible to generate several US slices from the signal recorded with a volumetric transducer (e.g., one 2D image per array row), thus a live 3D stream of images is created Figure 2.1 illustrates how a transducer for 3D array scanning systems is built. Several approaches are presented and evaluated by Yen *et al.* [48].

Volumetric phase array transducers binds maximum US volume resolution to the number of matrix elements in the transducer. Therefore, to get an image of reasonable dimension the volumetric phase array transducers needs to have an array of elements of reasonable size, hence these transducers themselves are on average large. Also, these systems are usually expensive. Because of the aforementioned reasons, they are not suited for liver intraoperative scanning.

2.1.4 Mechanical

Another way of creating volumetric US images is based on tomographic reconstruction of a conventional 2D US image series into an US volume. This could be done by utilization of a mechanical

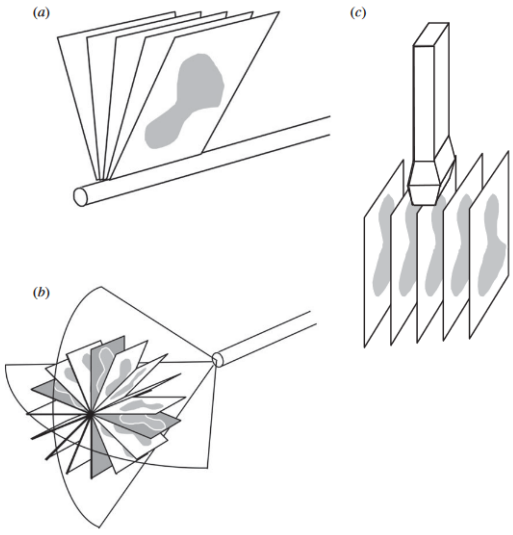


Figure 2.6: The figure shows three examples of mechanical scanning systems. (a) The transducer is rotated and scans the ROI in a radial way. (b) The transducer is rotated along its vertical axis and generates a "conical" volume. (c) A linear scanning can only move the transducer on a line. Source [11]

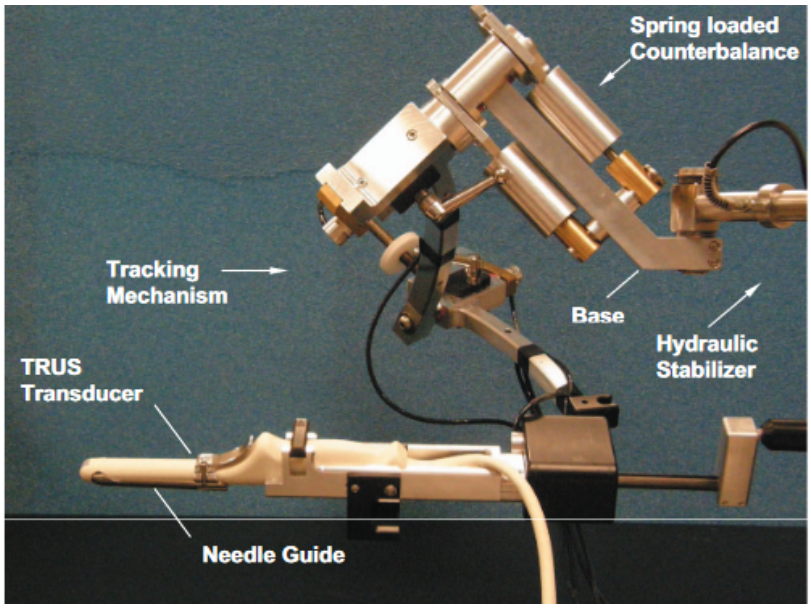


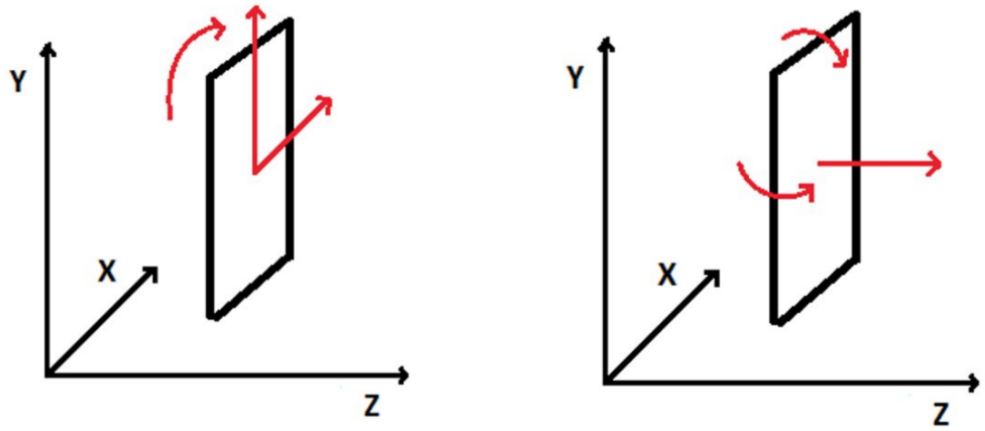
Figure 2.7: Example of a machine used for automated ultrasound guided prostate biopsy. Source [2]

scanning system (e.g., calibrated robot arm), which help to identify geometric constrains of each 2D US images and its subsequent placement into the 3D volume.

Mechanical scanning systems employ an ultrasound transducer that is attached to a motor-driven moving part, which will scan the ROI on the patient. Figure 2.6 shows three examples of mechanical scanning systems. The advantage of these systems is that the transducer will follow a continuous movement and generate US images at an homogeneous distance. However, they present some major downsides. Not only are these systems usually expensive but also they are not portable. Moreover, the design of a machine is often only suited for a small variety of US images. Either the kind of movement the machine can do, or the implementation itself, results in a tool meant for a very specific use-case. For example, the system pictured in Figure 2.7 is a design of a machine for ultrasound scanning in prostate biopsies. Therefore, they are only suited for certain intraoperative scanning procedures but for liver surgery being also large. Since the aim of this study is to help building a portable system for CAI, this class of tracking devices is not suited.

2.2 Reconstruction from Untracked Freehand US Sweeps

Ultrasound reconstruction is the process of aligning a sequence of 2D US images (also referred as stack or sweep) in three-dimensional respect to each-other. While performing a sweep, it can happen that the physician will not move the transducer in a linear motion, or will keep the transducer on the same area but rotate it around its vertical axis. Therefore, this procedure aims at translating and rotating the US images to their correct position in a 3D space.



ht

(a) In-plane motions and rotations

(b) Out-of-plane motions and rotations.

Figure 2.8: Schematic explanation of different kinds of motion in tomographic reconstruction from untracked freehand 2D US images.

The reconstruction process has to compensate for two different motions of the US transducer: in-plane and out-of-plane, also referred as elevational displacement [32]. In-plane motion is a translation or rotation on the US image plane, which can be successfully correct by several methods ([42], [30], [35]). The most challenging motion to correct is the elevational displacement. This is the motion caused by the rotation around the US image plane, which result in a drastic change in the structures present in the US images. It could also be caused by a sudden change in the speed of the US sweep.

Given that the transducer is held by a surgeon, this is moved with nonuniform speed and trajectory resulting in an uneven sampling of the ROI. Therefore, the data acquired with an US sweep is sparse and it is not known of what part of the ROI the US images are from.

Tomographic reconstruction from sparse data is a challenge that has been not only addressed in the field of US reconstruction but also in other medical imaging tasks. The FastMRI is a yearly challenge where various teams compete to get the best reconstructions (in terms of speed and accuracy) from MRI scans. The challenges in the 2019 edition were aiming at reconstructing a volume from one fourth or one eight of the data randomly picked from MRI scans. Particularly interesting for this study is the work from Pezzotti *et al.*. [31] that won the challenge of reconstruction from one eight of the data in the 2019 edition. They implemented a deep-learning based method for tomographic reconstruction and augmented its input with geometry knowledge of the accuracy of the scanner used to sample the dataset. They added information about the image shape specific to the MRI scanner used to sample the data in their dataset. The addition of these information resulted to positively influence the reconstruction, thus the idea of using transducer-specific geometry attention map for tomographic reconstruction from US images of the liver.

In addition to these challenges, the deformability and mobility of organs like the liver, make the US reconstruction process even harder. The deformability of the liver can change the look of the

anatomy in the US images depending on the amount of pressure used during the sweep.

The research in the field of tomographic reconstruction from 2D US data is split in two different main approaches. The first, bases the reconstruction process on pixel-based metrics the second on feature-based metrics. Only during the last years, US reconstruction methods based on machine-learning and deep-learning started being developed after their application were successful in other medical imaging fields.

2.2.1 Pixel-Based Reconstruction

The first category of algorithm is based on metrics that directly compare the values of the single pixels. Viola *et al.*[42] propose a method based on Mutual Information, an algorithm that tries to align the images maximizing the area in which they have same pixel values. First, they compute image entropy distribution with the Parzen Window method [10] and joint entropy. Then the computed entropy is used to compare samples from the images as a measure of complexity. The output is an estimation of a transformation matrix for which the entropy is the maximal and joint entropy will be minimal. In other terms, the algorithm will compute an estimation of the transformation matrix so that the information in common between two 2D US images is maximized. The algorithm manages to well correct in-plane motion, while does not compensate for out-of-plane motion.

A different pixel-based approach has been proposed by Oye *et al.* ([30]). The authors presented a reconstruction method based on the summed square distances of the pixels of neighboring US images. Their work presents some interesting points. First, instead of finding the transformation matrix only between two consecutive US images, they perform reconstruction on n consecutive 2D US images. They states that pairwise reconstruction would propagate the error from the previous steps. Pairwise reconstruction is done finding the transformation between a US images and its previous (step i), then this matrix is multiplied by the previous transformation (found at step $i-1$). Error propagation that would end up generating bad reconstruction towards the end of the US image. They instead compute the transformation matrix over n US images, computing the final transformation matrix as the squared sum of n pairwise transformations. This approach manages to better correct out-of-plane motion.

Another kind of pixel-based approach is based on Normalized Cross-Correlation (NCC) [14] between functions. These class of algorithms analyze the value of functions over the images (for example gamma function) to perform reconstruction. Algorithms that implements the NCC measure proved to work best on images with similar functions. Therefore, in their work on reconstruction of US of human bones [34] Schers *et al.*showed that using single pixel values and average pixel values from two 2D US images it is possible to achieve good performance in US reconstruction. However, this worked because bones have a very specific shape and are clearly visible in US images. Therefore, given that liver does not has such well defined anatomies, this approach is not suited for tomographic reconstruction from 2D US images of the liver.

2.2.2 Feature-Based Reconstruction

The second group uses feature-based metrics to perform tomographic reconstruction from 2D US images. Therefore, these class of algorithms adds a step to the reconstruction procedure. First, features are computed, then these are used to compare the images and place them in 3D space.

Several studies ([35][21]) implement SIFT to find scale invariant features for their reconstruction method. The method proposed by Schneider *et al.* [35], was developed for reconstruction of volumes with small displacement implementing a Register-to-Global strategy. First the method finds the SIFT-based descriptors. Then the newly found features are registered to a combination of previously found features from all the precedent US images. The features that will be stored for future matches are found first with a PW reconstruction. Then, the feature vectors from the pairwise reconstruction are analyzed to find symmetric matches, which will be then scanned with a RANSAC algorithm to remove outliers. The final transformation matrix will be computed with least-squares reconstruction algorithm from the matches that passed the RANSAC test.

Some of the approaches take advantage of speckle-based features. Speckle is a signal, which is is generated by constructive and destructive interference of the scattered ultrasound through different structure of the organs. It appears in the US images as a grainy pattern. Chen *et al.*[4] gave a first definition of the method called Speckle Decorrelation, which employs speckle from US images to include depth dependency when computing the transform matrix.

Following this idea, many studies tried to model and use speckle for tomographic reconstruction from 2D US data. Tao *et al.* [39] demonstrated that speckle can be satisfyingly approximate with a Gamma distribution. Wang *et al.* [43] instead used a Fisher-Tippet distribution to model speckle and then extract features for finding transform matrices. First, the method models two US images with the aforementioned distribution, then compares the two distributions with the symmetric Kullback-Liebler distance to find the transformation matrix. This method showed great results in scenarios where images have minimal differences and are almost only translated.

2.2.3 Statistical Based Reconstruction

During the last years, thanks to the evolution in the field of artificial intelligence, deep learning and machine learning methods have been implemented in medical imaging. In specific, the application of deep-learning based methods in tomographic reconstruction from sparse MRI/CT data and artefact reduction has shown great results.

During the latest years, tomographic reconstruction from sparse CT/CBCT/MRI data started to be the focus of many studies. The process of CT/CBCT/MRI scanning is long, therefore the idea of the research in this field is to reconstruct the images from a fraction of the scans. As a result, the scanning time will be also reduced to a fraction of the original. The FastMRI challenge, in its 2019 edition, proposed two different problems where it was asked to present a method that could reconstruct full MRI scans from or one fourth or one eight of the original data in the scan. Many deep-learning based approaches were proposed achieving great results in both challenges. Tomographic reconstruction from sparse CT/MRI data can be considered similar to tomographic

reconstruction from US data.

With the above considerations, it is clear why deep-learning based approaches have also started to be implemented in the field of US reconstruction.

In their work at [32], Prevost *et al.*proposed a Convolutional Neural Network model to predict motion in consequent US images. Their method is based on speckle decorrelation and optical flow prediction. It has been the very first approach that aimed at developing a reconstruction process that only relied on deep-learning There were other approaches which implemented deep-learning solutions, but these were or only replacing parts of the process, or improving results of analytical approaches, or removing unreliable parts from the estimation of the volume. In some cases, deep-learning based solutions have been implemented to use

On the other hand, CNNs do not require the definition of any feature to work. The definition of the features is in fact performed automatically by these networks, and will depend on how their training is performed. Therefore, the method proposed by Prevost *et al.* is a CNN based on FlowNet [9], that sees it adapted to work on US images and to output a six-dimensional vector.

The output of the model presented by Prevost *et al.* are 6 values representing the offset and rotation vectors of each 2D US image. The model takes as input two US images and optionally the optical flow computed between them as a two or four channel image (depending on the use of the optical flow) following the work from Dosovitskiy *et al.*, where it was proven that such procedure was giving better results that feeding the network two single channel images. The model described at [32] can also work with additional input data acquired from a Inertial Measurement unit (IMU). This was proved to help the prediction of depth displacement between images, however it is not suitable for liver intraoperative sweeps because of the additional space required by the IMU. However it will be not possible to use such data for this study. Besides the fact that the dataset used in this study was not built with an IMU, hence it has no inertial measurements, the implementation of such unit in an intraoperative US scanner is not ideal because of its dimensions (unless the transducer has an integrated IMU but there is none at the moment).

As it was previously discussed, the method from Prevost *et al.* is a reconstruction method based on speckle decorrelation. In Figure 2.9 it is shown a comparison between the reconstruction process using a CNN with the reconstruction using a speckle decorrelation implementation. From the comparison it is clear that the two processes are similar in terms of steps. The convolutional layers are thought to behave like a cross-correlation algorithm as described in Section 2.2.1, while the pooling layers are meant to work as patch-based approaches that sees the speckle decorrelation applied to patches of the images. The activation function in the CNN layers can be compared to the selection of relevant speckle in the speckle decorrelation approaches.

Even though the anatomy for which the method has been trained is much different then the one used in this study the approach of the model is supposed to work on any kind of anatomy or transducer ([32]). Although there is no work in literature that reports an adaptation of this method for deformable anatomies, the innovative approach proposed by Prevost *et al.* has good potential to be adapted to work for tomographic reconstruction from 2D US images of the liver.

Guo *et al.* [15] propose another CNN based method that uses speckle motion to predict US image

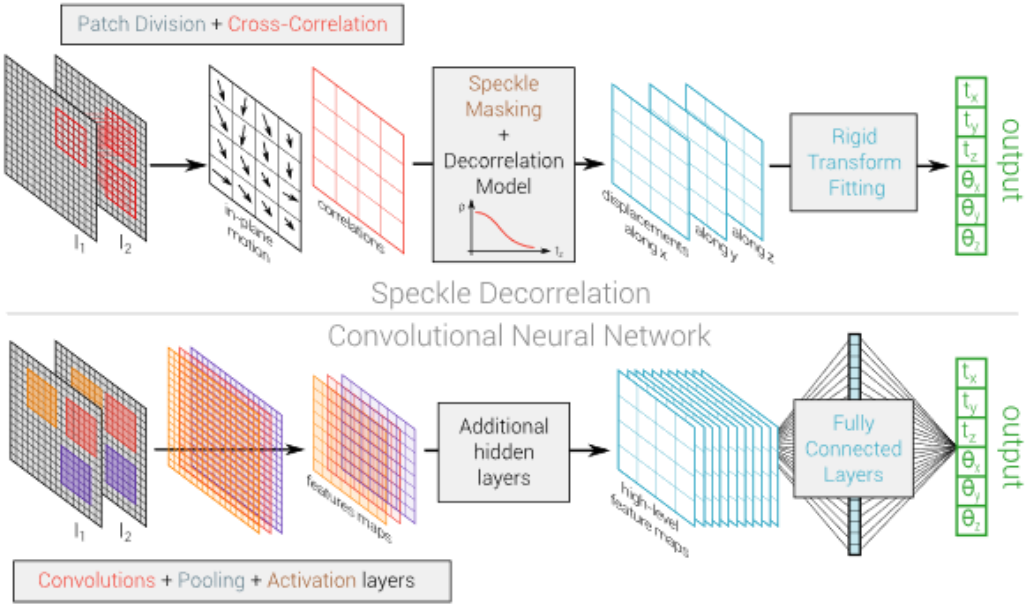


Figure 2.9: Similarities of the CNN proposed by Prevost *et al.* with previously presented speckle decorrelation based algorithms. Source [32].

positions. The model described is based on the ResNext architecture. A 3D convolution block takes as input a stack of US images (e.g., 5 is reported to be the best stack size). Its output will be the input of a residual layers block followed by a self-attention block. At last the model outputs 3 values for the offset and 3 for rotation of the US image. To be noted the use of 3D convolutions. The authors state that this was required to let the network better extract feature mappings along the axis of the channel (e.g., time). The network will focus on slight displacement of image features and thus be trained to connect these. Also, they introduced in the architecture a self-attention module to focus the learning of the network on speckle-rich areas. Therefore, the network will take as input the output of the residual block and output an attention map, which will be used to assign more weight to more relevant regions in the US image.

As it was previously discussed in section 2 the method proposed by Prevost *et. al.* [32] was the first were US reconstruction was entirely based on deep learning, bringing great innovation in this field of research.

2.3 Computation of Ultrasound Volumes

Many approaches to generate a 3D US volume from tracked or non-tracked scans have been implemented and assessed. The main challenges for this procedure is to avoid the generation of artifacts in the final volume and the loss of relevant data due to blurring.

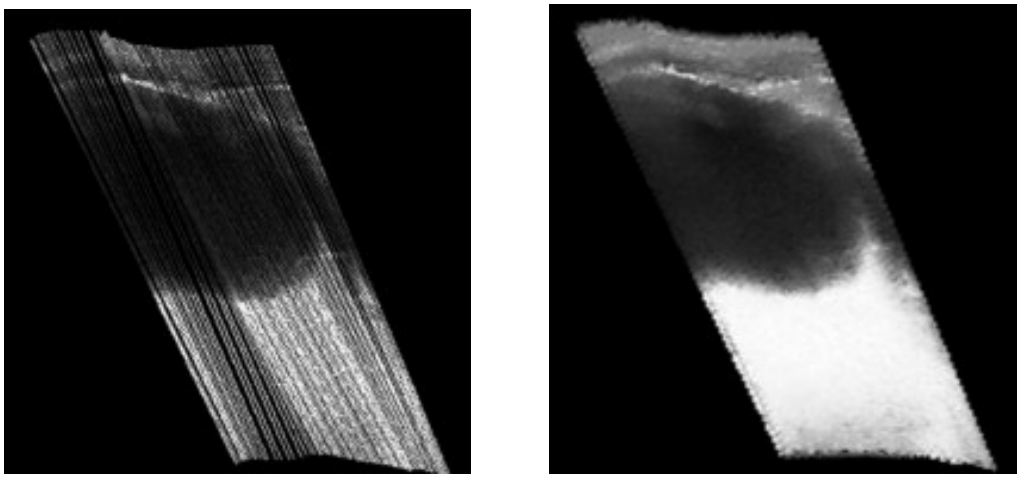
The methods reviewed for this study have been divided in four major categories summarized in the next sections.

2.3.1 Pixel-based Volume Computation

Pixel based approaches are methods that iterate over the pixels in the US images to decide what value to give to the voxels. Therefore, not all the voxels will necessarily be checked by an algorithm. All the following algorithms work for reconstruction of tracked US sweeps (e.g., the physical position of the US images is known thanks to a tracking system, or computed with a reconstruction algorithm, see previous Sec.).

An example of such algorithm is the Pixel Nearest Neighbor (PNN) [22]. This algorithm works in two steps, the bin-filling and hole-filling steps. In the first step, the algorithm assigns values to the voxels based on pixel values and US image positions. Iterating over the pixels in the US images, it will assign the value of a pixel to its nearest voxel. A voxel is considered the nearest based on the position and rotation of the US image in the coordinate system. This step will produce a volume with all the relevant data in the US image. However, depending on how sparse the US images are, the volume will present empty voxels. An example of a bin-filled volume is shown in Figure 2.10a. When this step is concluded, the algorithm will start the second step in which it will fill the holes. First it will have to identify which parts of the volume need to be filled, then it will chose the value for each filling voxel with, for example, the average of the voxels values in its neighborhood. Figure 2.10b shows an example of a reconstructed volume with a pixel-based method. A voxel is considered to be a neighbor depending on a chosen function.

In their work at [45] Wen *et al.*proposed an implementation of a PNN where this was determined by the radius of a sphere, while in the work at [13] it was a parallelepiped of different dimensions (the



(a) Example of volume after the bin-filling step (b) Example of volume after the hole-filling step

Figure 2.10: Example of the steps in the PNN algorithm. Source [46]

most relevant was a cube of side 2). In the study at [13], was evident that, using an average function to determine the empty voxel value, the adoption of larger neighborhoods result in blurrier hole-filled volumes. Not only the volumes generated from these methods usually result being blurred, but also they are prone at generating artifacts where the bill-filled volume meets the hole-filled. Therefore, further research have been done to improve this algorithm, especially in the hole-filling step.

With The Fast Marching Method [46] (FMM), the authors proposed a new way to interpolate in the hole-filling step. Their idea was to change the order in which the empty voxels are filled. Most of the previous methods were utilizing a linear interpolation, meaning that the voxels were filled following their physical position in the volume. What they implemented instead, was a filling order based on the shape of the boundaries of the holes to fill. The propagation of the voxel values follow the direction of the normal to the closest boundary. The neighborhood function for this method was the radius of a sphere and the voxels values are compute as an average. The experimental results showed a great improvement at removing blur from the volume.

2.3.2 Voxel-based Volume Computation

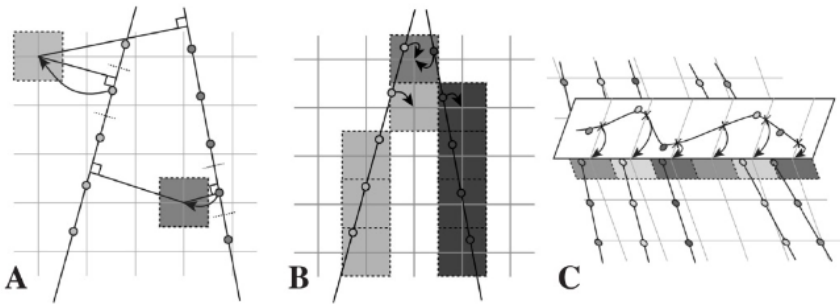


Figure 2.11: Volume computation methods. a) Voxel Nearest Neighbor algorithm. b) Distribution step for Pixel Nearest Neighbor. c) Functional Based Methods. Source [20]

Voxel based reconstruction methods try to answer the question "What data should be assigned to this voxel?". The general idea behind this class of algorithms is to iterate through all the voxels in the volume and compute a value from the 2D US images to assign to each voxel. This class of algorithms works with tracked US US images only.

The most basic of these algorithms is the Voxel Nearest Neighbor. As its name suggests, the algorithm travels over the voxels and assigns the physically nearest pixel value to the voxel. This procedure is good at preserving patterns from the 2D images, however it also preserves speckle and generates big artifacts where the data is more sparse.

Distance-weighted algorithms choose the value for each voxel based on the average value of the inverse weighted distance of the pixels in the neighborhood. First, for each voxel the value of the neighboring pixels is weighted based on the distance from the original. Then the average of the assigned values is given to the voxel. These algorithms are capable of suppressing speckle noise,

however they tend to blur the image because of the averaging step.

An improvement on distance-weighted algorithms was proposed by Huang *et al.*[16]. Their adaptive squared-weighted-distance algorithm uses kernels of adaptive dimension to decide which pixels are part of a neighborhood. The kernel dimension is based on the radius of a sphere. If the variance of a neighborhood is small, then the voxel is assigned the average value of the neighborhood as in the normal distance-weighted algorithm. If instead the variance is big, the kernel for neighborhood selection is set to be smaller until the variance becomes small. This approach solves part of the blurring problems in the final volume thanks to the adaptive kernel. The value for voxels on the edges will be averaged on a smaller number of non-relevant pixels. Also, it was proved to be slightly better in terms of computation.

Kernel regression algorithms are a different approach to Voxel-based reconstruction methods. Although they also presents a bin-filling stage, the second step, called regression stage, computes the value for the empty voxels with a weighting function. All symmetrical function can be used and one of the most popular ones is the Gaussian function as in [5]. The weights are meant to penalize more distant voxel values so they will have less impact on the value computation. This class of algorithms manages to well suppress the speckle noise without generating too much blur in the reconstructed volumes. Although this approach is more similar to a pixel-based approach [26], the regression step requires all the voxels to be checked, hence their computation duty is comparable to voxel-based algorithms.

A further improvement to kernel regression is the implementation of adaptive kernel dimension [18]. In this algorithm the dimension of the kernel used to assign voxel values change in relation to the density of the data. If the voxel is situated in a homogeneous neighborhood, then its value is computed as above. If instead the voxel is in a data-sparse area, the kernel dimension is diminished until its neighborhood is considered homogeneous. A neighborhood is considered homogeneous if the variance of the voxel values is smaller than a chosen parameter. Thanks to this improvement the algorithm is better capable at suppressing speckle noise without losing relevant data (e.g., sharpness).

In the latest years, with the development of GPGPU programming, some implementations of function-based reconstruction algorithms have significantly improved their running times. The works at [44] [6] are examples of such implementations.

2.3.3 Function-based Volume Computation

The algorithms that belong to this category are those that take as input an array of 2D reconstructed US images (e.g., a "video" recording), then base their computation of the value to assign to each voxel on a function and output a 3D US reconstruction. Even though the quality of the reconstructions generated with these methods is superior, they are computationally heavy, hence the interest for this field of research is low.

An example of function-based approach is the Bayesian-based method at [45]. The method aims at reducing loss of data on relevant edges in US images. As it was written in Section 2.3.1, the Fast Marching Method tried to remove blurring of relevant edges with good results. The aim of

this work was to further improve such loss. Therefore, the authors focused on a method capable of better assigning values to on-relevant-edge voxels. The idea was to use an adaptation of the Nonlocal Model (NLM) filter proposed at [3], which was originally designed to work for Gaussian noise removal. However, as it was previously discussed, speckle can be well modeled with a Gamma distribution. To adapt the NLM filter to work with this distribution, they rewrote the algorithm under a Bayesian formulation. The values of on-relevant-edges voxels are computed using such formula. The rest of non on-relevant-edge voxels are filled with the FMM algorithm.

Although the good results obtainable with the implementation of Function-based reconstruction methods, these methods are too computationally heavy to be used for this study.

Chapter 3

Materials and Methods

In this study we propose a method for tomographic reconstruction of liver US images, using freehand non-tracked ultrasound data and a deep learning CNN model with a transducer-specific geometry attention (TSGA) map. The performance of the proposed models, is evaluated using geometric calibration phantom and clinical patient data.

Given the research questions stated in Sec. 1.1 and the related work mentioned in Sec. 2, the aim of this study was to: (i) understand if the adaptation (the base model for this study) of the method proposed by Prevost *et al.* [32] for US imaging of body extremities can generate comparable results for the reconstruction of intraoperative US images. For the second research question, the study aims to explore if the performance of the model could be improved (ii) by adding a transducer-specific geometry attention map (e.g., incorporation of the information on 3D geometry of 2D US slab) or by (iii) considering the implementation multi-task training. In this study, 5 deep-learning models are proposed to estimate the rotational and translational displacement between two US images, subsequently used as direct input for reconstruction of a US volume.

In this project, three-dimensional displacement and rotation between two subsequent 2D US images is estimated using deep learning models. Thereafter, we suggest to utilize this information as the primarily input information for the tomographic reconstruction method. The suggested model takes inputs of two neighboring US images and predicts the displacement in spatio-rotational translation matrix between them. Once relative displacements of all neighboring images in a given US series have been determined, a volume can be reconstructed with a PNN algorithm ([37]), as implemented within the open source image-guidance platform CustusX [1].

Subsequently, two variations of the aforementioned base model are presented. The first variation of the model is the augmentation of its input with a transducer-specific geometry attention map. This attention map is incorporated in the form of a gradient map augmenting the input image, and is inversely proportional to the thickness of the US slab (see Fig. 3.1). Pixel intensity of the TSGA map is scaled to 0-255 range and are masked with a shape of the US image (e.g., cone, see Fig. 3.9). As it was discussed in Section 2.1.1, US beam of intraoperative transducers used in this study have a fixed focus depth. Therefore, actual resolution of the 2D US image varies along the Y axes

(height) of the image. Consequently, higher resolution areas of the images will be more sensitive to subtle motion (e.g., more relevant for training of the model), when compared to thicker parts of the slab. Hence, the TSGA map gives the higher importance (e.g., higher pixel intensity) to high resolution areas. We expect that transducer-specific geometry knowledge is expected to facilitate more accurate prediction of the out-of-plane displacement/rotations.

The second variation suggested in this work is an alternative method for optical flow incorporation within the CNN architecture. In this case, multi-task learning strategy is used to train the network for two tasks: predicting displacements between pairs of US images (pair-wise displacement) and estimating the optical flow between input images.

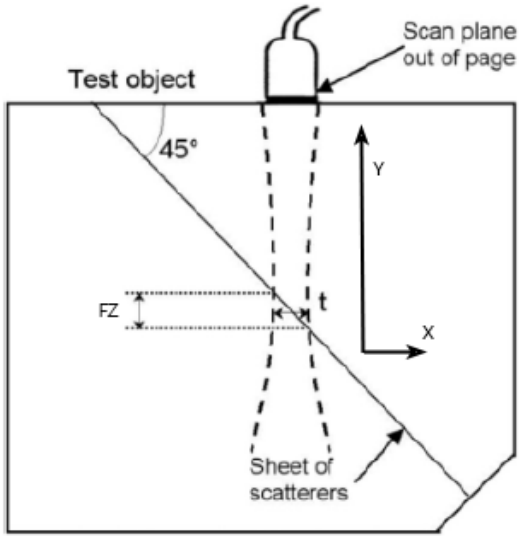


Figure 3.1: Visualization of the variation in thickness of the 2D US image along the Y axes.

The following sections will describe these methods in more details, starting with the description of the dataset used in the study. Then, the main hypotheses / challenges in the adaptation of the work of Prevost *et al.* will be assessed. The presentation of the two variations of the base model will conclude the chapter.

3.1 Dataset

Two types of data are used in this study: US images of geometric calibration phantoms and an intraoperative patient dataset. The phantom dataset consists of a single US sweep of the CIRS ATS model 539 US phantom ([7] manufactured by Computerized Imaging Reference Systems, Norfolk, USA). The patient dataset consists of 234 intraoperative US sweeps of the liver, acquired from 44 patients. All clinical data used in this study was acquired after getting an informed consent of a patient and is a part of the larger clinical study, reviewed and approved by the Medical Ethical Review

Committee of the Netherlands Cancer Institute Antoni van Leeuwenhoek Hospital (NL65724.031.18 / N18ULN number in the Dutch trial register).

In the patient dataset, nearly 65 thousands two-dimensional US 2D images are used. The sweeps were recorded using the Aurora EM-tracking system ([27] produced by Northern Digital Inc., Waterloo, Canada) and an intraoperative T-Shaped convex ultrasound transducer (I14C5T model [23] manufactured by BK Medical ApS, Peabody, USA). Clinical frequency range of 14 to 5 MHz and a focal range of 10-80 mm was used. Within a sweep, each separate 2D US image is saved as a grey-scale single-channel image with size of 544x668 pixels, and subsequently reconstructed into a 3D volume using the PNN algorithm and the EM-tracking information. Next, generic image quality of the reconstruction is scored by three experienced users. For each of the reconstructed US volumes, the severity of image artifacts, including pressure related deformation of the organ (e.g., bumpiness of the volume), as well linearity of transducer motion (e.g., straightness of the volume) were evaluated. As a results, image quality (IQ) of each reconstructed volume is classified into three categories: good, average and not acceptable. Only the first two IQ categories are included in the training dataset.

As an example, two US volumes are illustrated in Figure 3.2. The US volume illustrated in Fig. 3.2a is nearly straight, meaning that the US sweep was straight and continuous. On the other hand, the red arrow in Figure 3.2b illustrates the non-linearity of the US sweep trajectory (e.g., due to the pressure), resulting in a curved volume.

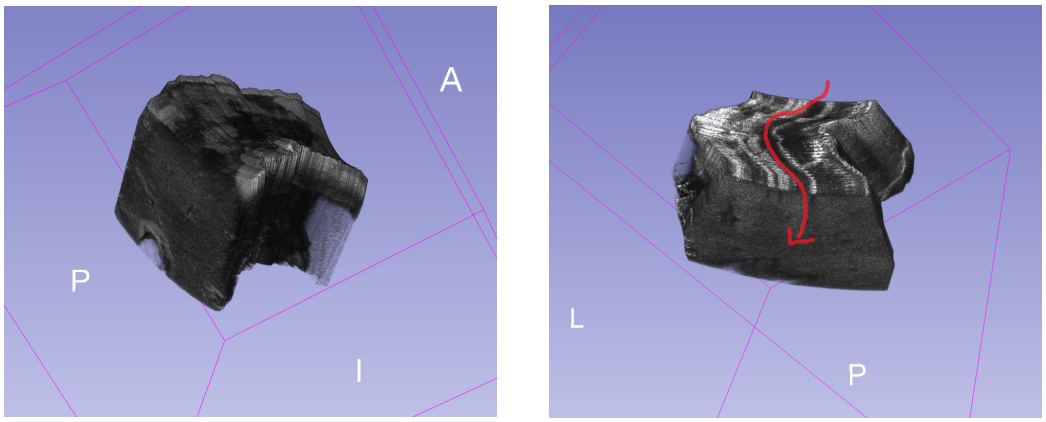


Figure 3.2: Examples of different ultrasound volumes from the dataset, an example of good volume in (a) and an example of an average volume in (b).

The ground truth used in the training and test sets are the US images in the sweeps with their relative spatio-rotational translation matrix obtained with the EM tracking system. In order to maximize the amount of the training data, the training dataset is augmented using two exploits. The first, being the horizontal flip of the 2D US images (ground truth transform matrix was changed accordingly). In this case, vertical flipping was not used, because it does not represent a realistic clinical situation (e.g., US transducer always project US images in the same way). As it was pointed

out in Section 2.1.4, the acoustic beam emitted from the US transducer is shaped symmetrically along the horizontal plane, but not along the vertical one. The second exploit is to pair together non-consecutive images in the original US sweep (e.g., to effectively enlarge spatio-temporal gap between the images). This not only helps to enlarging the final dimension of the dataset, but also helps the network to learn predictions of larger displacement between images ([32]). As it was described in Section 2.1.4, the clinician directly operates the transducer during freehand US imaging. Thus, large intra- and inter-operator variability in the speed and trajectory of US acquisition of the same ROIs is unavoidable. Effectively, this results in uneven sampling of the underlying anatomy.

However, the pairing has to follow certain constraints to avoid penalizing the training. It is not useful to pair images that are too far away from each other in the original sweep. In fact, it was proved to be counterproductive [32]. Therefore, the distance between two images that could be paired together will not exceed the gap of 5 images in the original sweep.

Next, the dataset is divided in a train and a test dataset. Table 3.1 indicates how the US images are assigned to the two sets. All images within a sweep are assigned to the same dataset, either test or train. This was necessary to reconstruct the full sweeps in the test set for qualitative analysis and evaluation.

		Total	Bad	Average	Not Acceptable
Train	Percentage	80	15	55	30
	Sweeps	176	26	97	53
	Pairs (thousands)	144	21.6	79.2	43.2
Test	Percentage	20	15	60	25
	Sweeps	46	7	27	12
	Pairs (thousands)	14	2.2	8.6	3.1

Table 3.1: Train and test set distribution in percentage, number of volumes and pairs of US images. Numbers of pairs are approximated and intended after augmentation.

The volumes are positioned in the reference coordinate space, as defined by the EM-tracking system. It is required to extract the volumes from this space, meaning that it is necessary to only take what is inside the oriented bounding box of the model. Then, the volume is translated and rotated so that its first image is in the origin of the reference space with no rotations. PCA is used for this task as it is widely adopted in literature to orient point cloud volumes. This step is required to have the volume in the same coordinate system as the registered stack of US images.

3.2 Model Adaptation

The following sections explains the adaptation process of the model proposed by Prevost *et al.* for body extremities on mobile and deformable anatomies. First the architecture of the base model is presented followed by the preprocessing used on the images in the dataset described in the previous Section 3.1. Then, the main challenges in the adaptation are explained and the solutions adopted

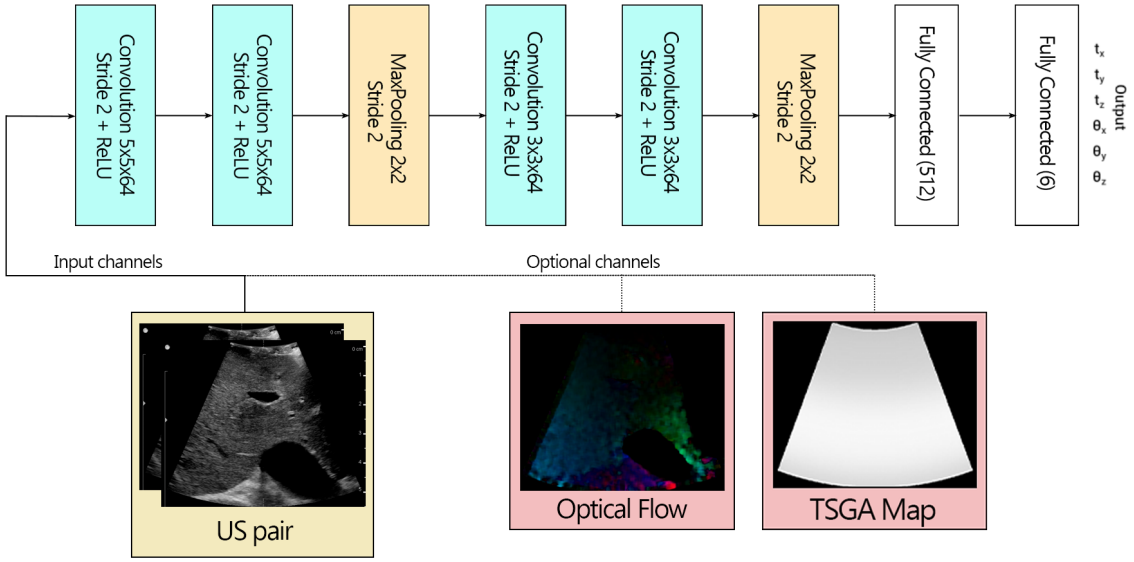


Figure 3.3: Architecture of the base model.

in this study are described.

3.2.1 Base Model - Architecture

The architecture of the base model used in this study is the one illustrated in Fig. 3.3. The model is a Convolutional Neural Network with a total of 8 layers. It is implemented in Python using the Tensorflow and Keras libraries.

As it was mentioned in Sec. 2.3.3, the architecture is an adaptation of FlowNet2 [9] to the applications on US imaging and predict a 6-dimensional displacement vector as output. The network's architecture follows the one of a standard CNN, where two convolutional layers are followed by a maxpooling and a dropout layers. This scheme is repeated twice and then followed by a fully connected layer, another dropout and the final fully connected layer with output of a 6 dimensional vector. The convolutional layers are all activated with a ReLU activation function. The ratio used in the dropout layers is 0.25, where the dropout layer was required to avoid overfitting.

The input of the CNN is a 4-channel image where the first two channels are the two consequent US images and the other two are the optical flow in the form of a bi-dimensional vector field. The optical flow is computed with the Gunnar-Farneback approach, as it was done by Prevost *et al.*. The channels of the input are normalized to range from 0 to 255. The number of filters applied with the convolutions, the strides and the maxpooling layers are adapted from the model of Prevost *et al.* since the input image size used in their work is similar to the one used in this study. The output of this models is a 6-dimensional vector containing three values (one for each dimension in space) for the translation displacement and three for the rotation displacement.

Concatenating the neighboring images as channels of input is inspired by the work of Fisher *et al.* [9], where they proposed FlowNet, a CNN model for optical flow estimation between 2D images. The CNN takes an input with two frames as channels and outputs the predicted optical flow between them. Here, the authors not only presented the model, but also they studied what was the best way to feed a pair of frames to their model. It resulted that concatenating them in a single image was the better choice. Thus, the same approach is used in this study. Moreover, FlowNet achieved good performance at predicting optical flow on several datasets. Given the great results showed by FlowNet on optical-flow estimation and the good results achieved by Prevost *et al.* adapting that architecture to perform tomographic reconstruction from 2D US images, this architecture will be also implemented to work on the dataset described in Sec. 3.1.

3.2.2 Preprocessing

The method proposed by Prevost *et al.* is based on the idea of speckle decorrelation, which means that the prediction of displacement between consequent images is computed on how the speckle moved across the images. Speckle is a low-level feature in the US images and, as a consequence, they did not apply any noise-removal filter to the US images. To support this decision, Prevost *et al.* experimented training their model by applying different noise-removal filters to the US images and trained their model with the filtered images. They made a comparison of the results obtained by the model trained with filtered vs. unfiltered images, showing that the model trained not using any filter on the US images performed better.

Therefore, no noise-removal filters nor downsampling is applied to the images in the dataset described in Sec. 3.1 to train the models presented in this study.

3.2.3 Adaptation Challenges

The main challenges in the adaptation of the the work of Prevost *et al.* lie in the clinical application and the anatomy-related challenges of volumetric ultrasound data. The following factors add further difficulties to the adaptation process:

1. **The anatomy to be reconstructed**. While the dataset from Prevost *et al.* primarily contained images of upper extremities (e.g., little to no pressure-related deformation), the current patient dataset is solely based on intraoperative images of the liver. Unlike the extremities, the liver is highly mobile and deformable, especially after the laparotomy. Moreover, it is characterized by high inter-patient variability (about 40%). During intraoperative imaging, the liver is dynamically changing its shape, depending on the amount of pressure used by the physician. As a results, the acquired scan of the liver is non-uniformly deformed in the direction of the acquisition, due to associated organ deformation and the pressure. This results in additional challenges for the reconstruction models. The models will have to handle sudden changes in the look of the anatomies in the scans, which can compromise the final result.

2. **The dataset dimensions**. The dataset used by Prevost *et al.* contains more tracked volumes than the one used for this study. The dataset used by Prevost *et al.* consists in 700 US volumes and a total of 355 thousands 2D US images, which were subsequently augmented with horizontal flipping and pairing of non-consectutive scans. Since Prevost *et al.* states in [32] that their method can be better generalized to other anatomies (in their case they generalized from reconstructing sweeps of extremities to sweeps of the carotid) when using a model trained on more data (e.g., pretrained on US images of extremities then finalized on US images of the carotid), having a smaller dataset adds a further challenge for the presented models.

3. **The ultrasound transducer used**. The transducer used in the work from Prevost *et al.* is a percutaneal linear probe, while the images in this dataset are acquired with a conical intraoperative scanner. As it was pointed out in section 2, the images generated from a convex transducer impose slide degree of deformation to the organ (e.g., due to its shape). Additionally, unlike linear transducers, they have a "dead zone" near direct contact with the transducer, thus do not provide meaningful image information in the near FOV. It is therefore required to exclude this irrelevant information prior to the training of the model

4. **The tracking system.**The dataset used in this study has been built with an electromagnetic-tracked ultrasound probe, while the one by Prevost *et al.* was generated with an optical tracking system. The accuracy of the optical tracking system is higher, having an average tracking error of 0.25 millimeters, against an error of 0.5 millimeters of the electromagnetic one (these errors are stated by the tracking system producer). As a result, the ground truth, used for the training of the models in this study is possibly more affected by tracking errors, which could negatively affect the performance.

3.2.4 Adaptation Solutions

Although Prevost *et al.* already illustrated that it is possible to implement their method for imaging of other anatomies, its performances was not evaluated for intraoperative imaging scans of the liver. Consequently, this was translated into the first research sub-Question 1 of this study.

In this work, a decay value of 0.9 to the learning rate is applied in the training of the presented models, to deal with the difference in dataset dimensions. In Prevost *et al.*, the authors did not apply any decay to the learning rate, yet, since the dataset described in Sec. 3.1 is smaller, it was decided to apply the decay to avoid overfitting during the training.

The discrepancy in accuracy of tracking systems used in Prevost et al. and the current study has no immediate solution. Therefore, no action to compensate for the difference in tracking system has been used in the adaptation process.

To assess the third adaptation challenge, two options are taken into consideration to handle with the the different US transducer used, which are:

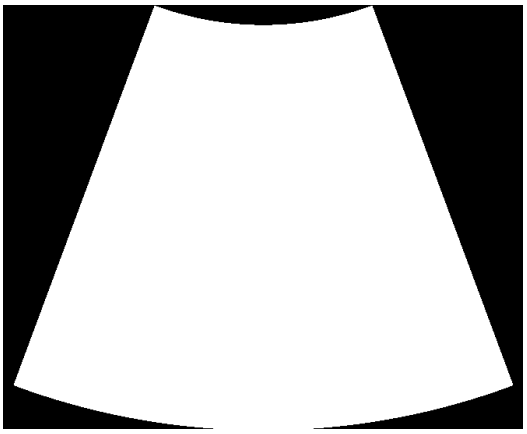


Figure 3.4: Example of attention map used to train the Base-Soft model.

- **Base-Soft**. This approach is called Base-Soft since it is based on the concept of soft attention. In this case, the application of soft attention was performed concatenating a binary attention map with the convex shape of the US cone as an additional channel of the input. Using the full US image does not remove any relevant data from the images, however the irrelevant frame around the US cone is not removed. The mask is meant to make the model focus only on relevant information and avoiding its training to be affected by the irrelevant frame around the US cone in the image. An example of attention map used to train the model is illustrated in Fig. 3.4.

- **Base-Hard**. This approach was called after the concept of hard attention. The hard attention concept here is implemented square-cropping the 2D US images in the dataset from the center to only contain relevant information. Even though this solution implies loss of relevant regions, it is a easiest way to reproduce images as similar as similar as possible to those in the dataset used by Prevost *et al.*. Moreover, because of what was discussed in Section 2.1.4 about how the US beam is shaped, the most relevant motion-wise information is in the center of the image. Consequently, the images are cropped from the center. Fig. 3.5b gives an example of cropped ultrasound image.

As a result, two models with different input sizes following the two aforementioned attention approaches are presented in this study. The input size of the first one is of 544x668x5 (Base-Soft), while the second one is 300x300x4 (Base-Hard). The first input size is chosen since it is the full size of images. This results in not wasting any relevant information from the images, leaving on the other hand an area of irrelevant data in the input image. An example of a full 2D US image from the dataset with irrelevant data around the US cone beam is illustrated by Figure 3.5a. To help the network focuses on relevant information, a binary mask is added as 5^{th} channel to the input image. Figure 3.4 shows an example of binary attention mask.

The input size of Base-Hard is a 300x300x4 image where 2 channels are the images and the other two are the corresponding optical flow. Although this results in wasting part of the relevant

information of the US images, all the irrelevant information are removed from the input. Moreover, in contrast with Base-soft, it is unnecessary to concatenate a binary attention mask to the input. The input image is a cropped square from the center of the original US image as in Fig. 3.5. The dimensions of the input image for this second model follows from what was discussed in Sec.3.2.4. Therefore, the biggest square that can be cropped from the center of the US images and that only contains relevant information has a side of 300 pixels.

(a) Example of US image from the dataset. The red square indicates the cropped area, a square with side of 300 pixels

(b) Example of cropped image used to train the Base-Hard model.

Figure 3.5: Ultrasound image cropping, from the original US image with dimension 544x668 an area of with size 300x300 is cropped from the center of the US image.

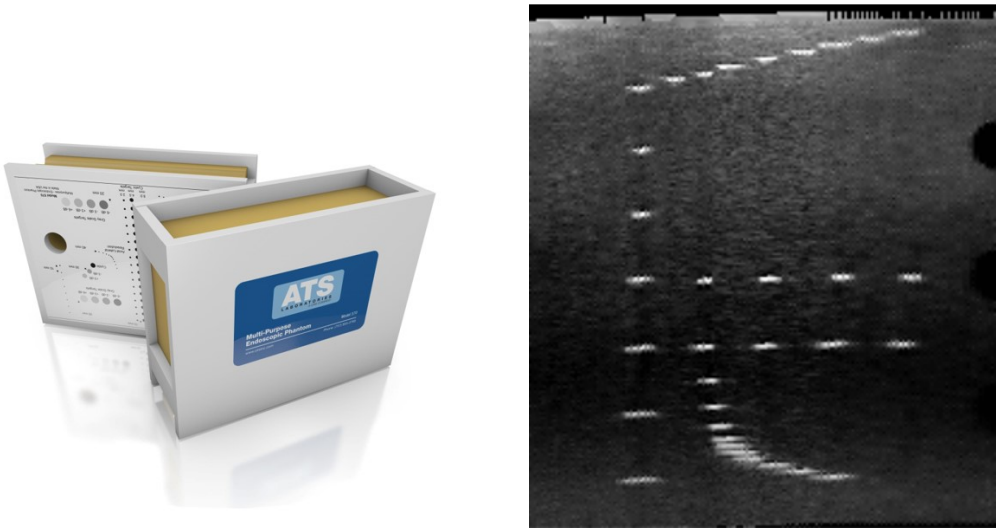
3.3 Model Variations

The following section presents two variations of the base model described in Section 3.2.1. The section starts with the integration of the TSGA map, to continue with the implementation of multi-task training on the base model.

3.3.1 Transducer-Specific Geometry Attention Map

This first variation of the base model described in Sec. 3.2.1 is an integration of additional knowledge related to the shape of the US slab generated by the US transducer into the model. Therefore, the input of the model described in Sec. 3.2.1 is focused with a transducer-specific geometry attention map, in order to exploit the apriori knowledge about the shape of the US image. It is an innovative approach in the field of tomographic reconstruction from US data and, to our knowledge, it has not been yet reported in the literature of this field. The addition of MRI-coil specific geometric sensitivity during training of CNN-based reconstructions has bee investigated before, and had beneficial effects

on geometric accuracy of reconstructed images([31]). Thus, since sparse MRI and US reconstruction are fundamentally similar challenges, in this study, it was decided to investigate if the addition of TSGA will be able to improve geometric accuracy of tomographic US reconstruction as well.



(a) The US phantom used to compute the attention map

(b) A slice from the EM-tracking based reconstructed volume of the area used to compute the TSGA map.

Figure 3.6: The US phantom and detail of the area used in the computation of the TSGA map.

The TSGA map is intended as a gradient map that will help the network to focus on the most relevant motion in the US images. This is added as additional channel to the input image in the form of a 0 to 255 single channel image. It was chosen to add the map as separate channel to follow the same approach that was used for the addition of optical flow to the input as it was done in [32].

The idea behind the computation of a TSGA map is based on the functioning of a US transducer. As it was discussed in Section 2.1.4, an array of US emitters is on the surface of contact and generates a beam with a certain shape that depends on the transducer. The beam has a focal area where the resolution is higher. Therefore, because of how the reconstruction method was modeled, the highest resolution areas will be the most relevant ones since the speckle motion will be the most accurate. The further the areas form the focal point the less relevant the motion will be. This is, again, because of how the image is reconstructed from the US waves. The intensity of each pixel is defined as the average value in the scanned area. Therefore, in the areas far from the focal point the pixel intensity will be given from an average of a bigger scanned portion, resulting in a value with less resolution.

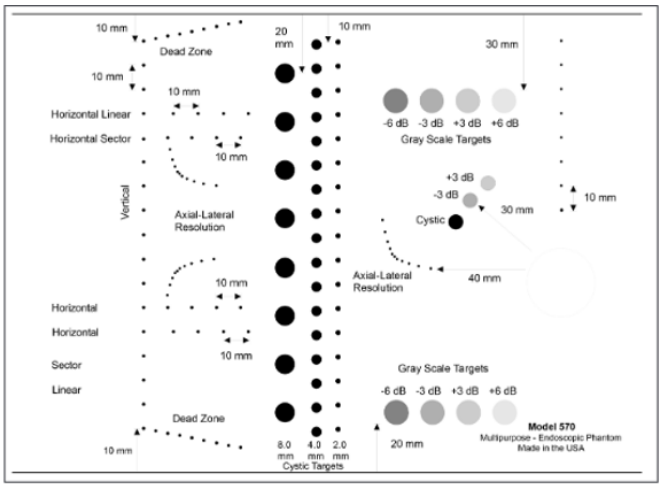


Figure 3.7: Gold standard of the US phantom used for the TSGA map computation.

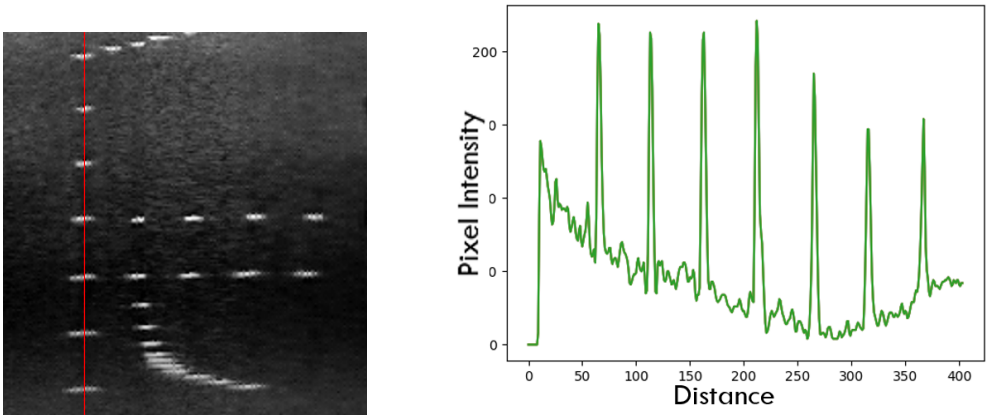
It is now clear why the addition of a TSGA map can positively influence the training of the proposed model. The model will better focus on areas with higher resolution, hence giving more importance to relevant speckle motion.

Computation:

The process of computing the TSGA map starts with the sweep of a US phantom (see Fig. 3.6a). The phantom contains beacons positioned at a standardized distance of 10mm in horizontal and vertical directions. The phantom used to compute the TSGA map is the CIRS model ATS 539. The portion of phantom that was scanned to compute the TSGA map is the one in Figure 3.6b.

The procedure used to compute the TSGA map is the following:

- The transducer was set to standard acquisition parameters and automatic focusing depth and focused on vertical beacon insert of the phantom (see Figure 3.6b, and on the left of Fig. 3.7);
- The area of interest of the phantom is scanned (Fig. 3.6b) with the same intraoperative transducer mentioned at 3.1 and images are stored;
- An imaginary line that cross the beacons vertically (Fig. 3.8a) is drawn;
- A pixel intensity profile is computed. Figure 3.8b shows a graph of the intensity value computed. The y axis will be the pixel intensity and the x axis will be the height in the US image;
- The computed distribution can be compared to known ground distribution intensity graph (Fig. 3.8b) to understand how different from reality the scanned image is;
- The Full Width at Half Maximum (FWHM) is computed for every peak in the two graphs, after what compared to the calibration width of the beacon. The half of this difference, being



(a) Example of imaginary line drawn to compute intensity profile of the beacons in the phantom sweep.

(b) Example of line pixel intensity line profile.

Figure 3.8: Procedure followed to compute intensity profiles required in the computation of the attention map.

absolute spreading of the image, is reflecting resolution degradation with respect to actual dimensions of the object. The lower the spreading the higher the resolution in the area of the image;

- The average spreading is calculated for each beacon illustrated in Fig. 3.8. The measurement was repeated to at least 3 separate image acquisitions, after what this values were converted to relative image resolution gradients. At last, TSGA was acquired by inverting the contrast of the resolution gradient (e.g., thinner the slab the higher the attention);

Figure 3.9 illustrates the output map, corresponding to a single channel image with values in the 0-255 range. The map has slightly higher intensity in the lower-mid area while the top and bottom parts have lower intensities (e.g., thicker slab). Thus, the intraoperative transducer has a higher resolution in the lower-mid area, where the pixel intensity of the map is higher.

The BK transducer used in this study has auto-focus set on the center of the US image, as a consequence it is possible to use the same map at every depth. If otherwise, it would be necessary to compute multiple maps depending on where the focus is set.

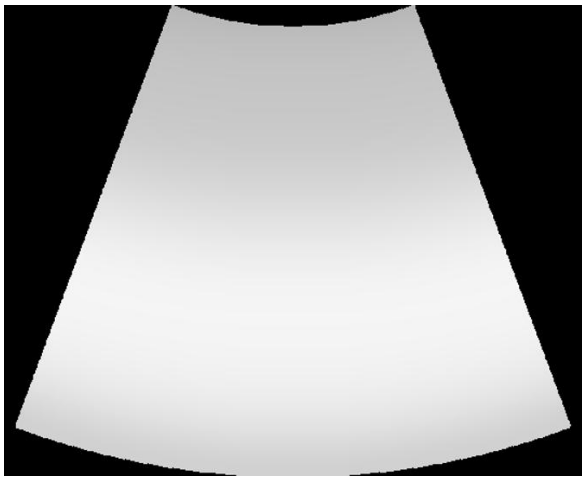


Figure 3.9: The TSGA map computed for the BK T-Shaped Intraoperative I14C5T transducer.

Usage: As it was mentioned above, the pixel intensities in the map span from 0 to 255, which is the same range of other the channels in the input image.

The TSGA map is integrated in the model concatenating it as 5^{th} channel of the input image. The map is cropped from the center according to the dimensions of the other channels. It is expected that, with the incorporation of TSGA in the model, the estimation of out-of-plane motion is improved.

Since, to our knowledge, there is no previous work in the literature of this field that implements TSGA for tomographic reconstruction from US 2D images, therefore it is known how the addition will affect the process, two models will be tested in this study: Attention-0.5 and Attention-1.0.
Attention-0.5 is trained with the TSGA map with half of its full weights, hence pixel values in the range from 0 to 127 (rounded down).
Attention-1.0 will be trained using the TSGA map with its full weights, hence pixel values in the range of 0 to 255.

3.3.2 Multi-Task Learning

This study proposes the implementation of multi-task learning to feed the optical flow information to the network in a different way than it was previously done in [32].

As it was discussed in Sec. 3, in their work, Prevost *et al.* decided to give the US images to estimate displacement of as a two channel image, in the same fashion proposed by FlowNet ([9]). In the same way, they augmented the model input with the optical flow adding two channel to the image with the two vector fields resulting in a 4 channel input image, same as described in Sec. 3.2.1.

As it was discussed in Section 2.2.3, the architecture of FlowNet has been implemented in a multi-task scenario successfully [28]. ActionFlowNet is a CNN for action detection where Yue *et al.* used the FlowNet architecture for the estimation of optical flow and a second branch to output a prediction of the class of the action in the image. Moreover, multi-task learning has shown promising

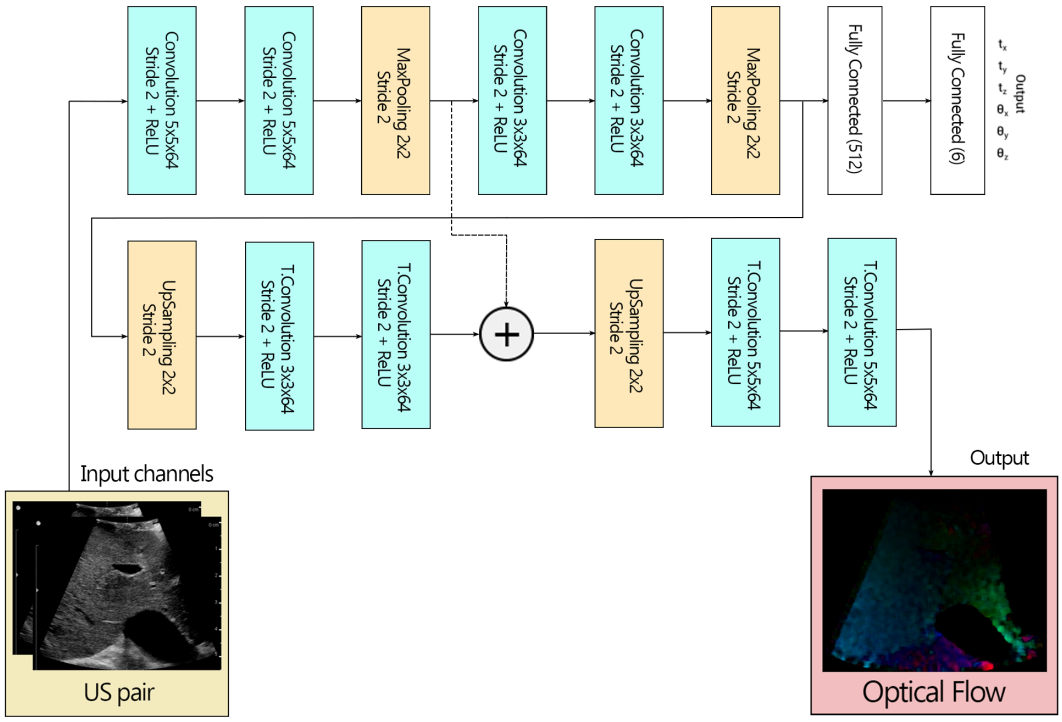


Figure 3.10: Architecture of the MultiTask model.

results in other medical imaging tasks like segmentation and classification ([12], [24]). Given the previous studies on optical flow estimation with multi-task learning, this study proposes MultiTask, a model that implements an alternative way to integrate the optical flow information in the model described in Sec.3.2.1. The optical flow information, instead of being fed to the network as additional channels to the input image, is instead used as ground truth for multi-task training of the base model to predict the optical flow and to estimate the motion between two input US images.

Thus, MultiTask is a variation of the model described in Sec 3.2.1 where it outputs not only the aforementioned six-dimensional vector, but also the optical flow between two input images. The training aims at minimizing the MSE loss for both tasks. The optical flow computed with the same algorithm as in 3.2.1 is used as ground truth for the training of optical flow prediction.

Figure 3.10 illustrates the architecture of MultiTask. The model have a first part (downsampling step) that coincides with the convolution and maxpooling layers from the base model described in Sec. 3.2.1. After the downsampling step, the MultiTask model splits in two:

- A first part of the model coincides with the two fully connected layers from the base model, and it is intended to generate the 6-dimensional vector as previously described.

- The other part of the model is a series of upscaling and transpose convolutions required to output a two-channel estimated optical flow of the same dimensions as the input image. The upscaling part of the network, the layers and their dimensions are inspired from the work of Fisher *et al.* [9]. Because of the lower number of layers in the network, a single skip from the down-sampling part of the model to the upscaling is used. The output of the first maxpooling layer in the downsampling stage is summed to the input of the first upscaling layer in the upsampling stage.

The weights in the downsampling step of the model are shared in the training for both tasks. The model is trained alternating the two tasks, and the MSE is used as loss function for both tasks.

Alternate training is used since the dimensions of the datasets and the loss function are the same for the two tasks. Although in some cases L1 loss is used in training for optical flow estimation, Fisher *et al.* trained FlowNet with MSE loss. Therefore, given that the optical flow estimation part of this study's model is derived from FlowNet, MSE is used in the training of both tasks.

Since the most suitable number of epochs to alternate between the two tasks is not trivial to find, a grid search is performed. The two variables in the grid search are the numbers of epochs for the two tasks, ranging from 1 to 25 with intermediate steps of 5. A number equal 15 epochs for each task resulted to be the best setup, hence used in the multi-task training.

Because of how convolutions work, it is necessary to adapt the dimensions of the input image for to the model output an image with the same shape as the input. As a result, while the input of the model described in the previous section is a 300x300x4 image, the input of this model is changed to be a smaller 256x256x2 image. With this input resolution, it is possible to use padding in the convolution layers to have symmetrical dimensions in the convolutions of the down- and upsampling steps. The output of MultiTask is the 6-dimensional vector previously described and a 256x256x2 image representing the estimated optical flow between the input US images. This shape is selected since is the closest one to the original 300x300x4, hence it is the best way to avoid wasting information from the US images.

Chapter 4

Experiments and Results

In this chapter, the evaluation and results of the presented models will be described.

The proposed methods were evaluated through two different experiments, the first aimed at giving a quantitative analysis of the reconstruction of the US phantom sweep, while the second will give a quantitative estimation of the reconstructions on patient data. The evaluation on the US phantom sweep was performed on the reconstruction obtained with the EM tracking data and the five presented models in relation to the gold standard of the US phantom. To evaluate the reconstruction models presented, 8 metrics were used in this experiment. The US phantom scanned was the ATS US phantom model 539, the same one described in Section 3.3.1. The reconstruction done with the EM tracking was added for measure of the accuracy of the tracking system.

The quantitative analysis on patient data was performed using the metrics used by Prevost *et al.* so that a direct comparison between the here presented models and theirs (as reported in [32]) was possible. To evaluate the reconstruction abilities of the presented models, 8 error metrics were computed on the tomographic reconstructions of 34 US volumes in the test set (see Tab. 3.1). The errors were computed in relation to the EM tracking based reconstructions. The metrics aim at understanding the models' performances in terms of in-plane and out-of-plane motion estimation.

The following sections will first describe the two experiments used to evaluate the presented models. Thereafter, the training and hyper-parameter optimization procedure of the reconstruction models presented will be explained. The results obtained with the two experiments will be presented and analyzed. It will follow a qualitative analysis of reconstructions obtained with the presented models. An analysis of the limitations of the proposed methods and conclusive thoughts will end the chapter.

4.1 Experiments and Evaluation Metrics

This section will describe the two experiments designed to evaluate the reconstruction abilities of the presented models. The evaluation metrics and the process followed to extract them will be also illustrated.

4.1.1 US Phantom Measure

This experiment was designed to compute eight metrics from the analysis of the reconstruction of the ATS US phantom Model 539. The phantom sweep was performed with the same US transducer that was used to build the dataset described in Section 3.1, the T-Shaped Intraoperative I14C5T produced by BK Medical. This transducer was chosen to be as coherent as possible to the data used to train the models. The sweep was tracked with the Aurora EM tracking system, again the same used to build the dataset. The tracking information were required to reconstruct the volume from the sweep and compare it to the other models. Moreover, from the analysis of the EM tracked reconstruction, it was possible to understand how reliable is the ground truth data used to train the presented models.

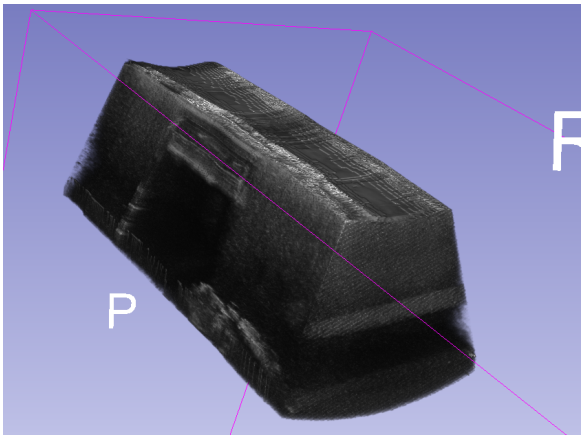
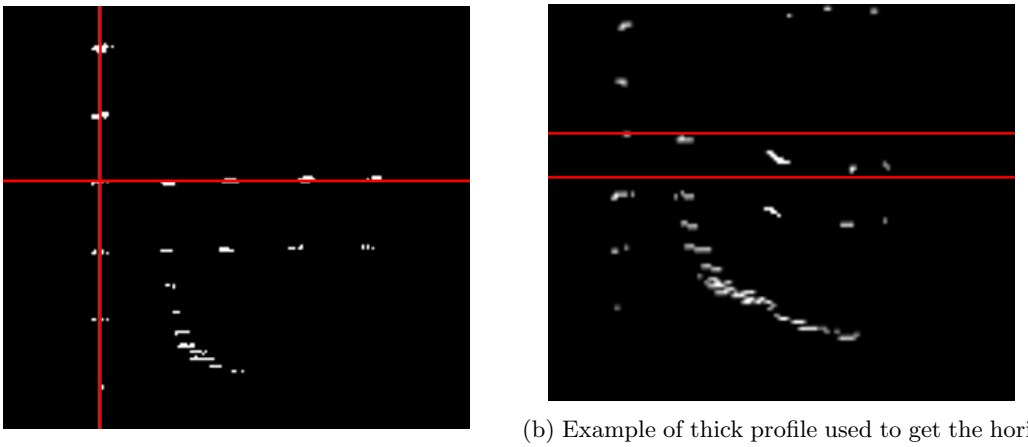


Figure 4.1: Reconstruction of the US phantom sweep with EM tracking data.

The phantom was scanned while positioned on the table and the EM tracked sweep stored in the DICOM format. To find the position of the beacons it was required to extract, process and analyze slices from the reconstructed volumes. Slices like the one in Figure 3.6b were extracted from the 3D visualization of the phantom sweep reconstructed with the EM tracking and the 5 CNN models.

To find the distances between the beacons, the slices were first processed (see Fig. 4.2). A threshold of 60% of the maximum value was applied to the slice to convert it to a binary image. Then, the region of interest was cropped from the slice and an intensity profile of variable thickness that goes across the beacons drawn. The thickness of the intensity profile could vary since some of the reconstructions did not place all the beacons on the same line. From the analysis of this intensity profile, the distances between the center points of the beacons in terms of pixel number was computed. The number of pixel was then multiplied by the pixel dimension stored by the scanning system during the sweep to get the intra-beacon distance in millimeters. This process was used for both horizontal and vertical intra-beacon distance computation. The errors of the reconstruction of the vertical line of beacons are related to in-plane motion estimation while those on the horizontal one are related to out-of-plane motion estimation.



(a) Example of pixel intensity profiles extracted from the horizontal and vertical lines of beacons.

(b) Example of thick profile used to get the horizontal intra-beacon distance from the Attention-0.5 reconstruction.

Figure 4.2: Pixel intensity line profiles for intra-beacon distance computation.

Four values are computed for both the estimated motions (total of 8):

- **Average error**. The average error is the mean error committed by the reconstruction methods at placing the beacons at the correct distance. Therefore, given that the gold standard of the intra-beacon distance was 10.0mm, if in a reconstruction all the beacons were placed at a distance of 9.5mm the average error was 0.5mm. The formula used is:

$$\frac{1}{n-1}\sum_{i=1}^{n-1}|10-l_i|$$

  where, l_i is the i-th gap between the beacons and n is the total number of beacons. The value will be the average error between the beacons in millimeters. This value helps to better understand how did a reconstruction method performed overall at estimating a in- or out-of-plane motion.

- **First and third interquartiles**. The first and third interquartiles are computed on the distribution of intra-beacon errors.

- **Absolute error**. The absolute error is the difference between ground truth and reconstructed distance of the first to the last beacon. This value indicates how local errors propagated throughout the reconstruction. Therefore, the value was computed as $|l_{gt} - l_p|$ where l_{gt} is the total ground truth length of the distance between first to last beacon and l_p is the one from the reconstructions.

4.1.2 Experiment on Patient Data

In Section 3.1, the dataset used in this study and how it was divided in a train and a test set was indicated. To perform this analysis, the volumes in the "Good" and "Average" categories of the test dataset were reconstructed with the proposed methods and the EM tracking. All the error metrics were computed in relation to the reconstruction of the sweeps based on EM tracking data.

Eight metrics are used to quantitatively evaluate the reconstruction methods, which are:

- **Average absolute parameter-wise error in a volume**. The average absolute parameter-wise error is a 6-dimensional vector containing the mean error for each parameter committed by the model when estimating displacement between the pairs in a sweep. The three translation parameters are referred as t_x, t_y and t_z while the three rotation parameters are referred as r_x, r_y, r_z. Therefore, the formula used to compute the absolute average error for each of the six parameter is

$$\frac{1}{N}\sum_{k=1}^{N}|\vec{p}_{i,k,e} - \vec{p}_{i,k,gt}|$$

  where k is a generic displacement vector between two neighbor US images in the sweep, N is the total number of pairs in the sweep and $\vec{p}_{k,e}$ is the i^{th} parameter of the k^{th} estimated displacement vector and $\vec{p}_{k,gt}$ is the i^{th} parameter of the k^{th} ground truth (e.g., acquired with EM-tracking) displacement vector in the sweep.

- **Drift**. This value is intended to measure the divergence of the estimated trajectory from the ground truth. It gives a measure to the propagation of the error of in-plane displacement estimation throughout the reconstruction. The value was computed by absolute difference between the central position of the first and the last US images between estimated and ground truth reconstructions. As a result, the rotation was not involved in this metric. Considering that the volumes were placed in the origin of the same reference space (see Sec. 3.1), the formula used to compute the drift is

$$\text{drift} = |\vec{t}_{n,p} - \vec{t}_{n,gt}|$$

  where, $\vec{t}_{n,p}$ is the position vector (e.g., 3-dimensional without rotation) of the last US image of the reconstructed volume in space and $\vec{t}_{n,gt}$ is the position of the last US image of the ground truth reconstruction. The value is an absolute distance in millimeters.

- **Length Error**. This value represents the difference in total length of the volume (first to last image position in terms of translation) with the ground truth. This error was required to understand the propagation of the error of out-of-plane motion estimation throughout the reconstruction.

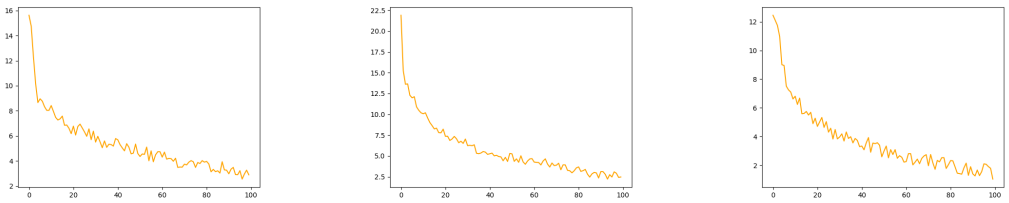
4.2 Training

The five models for tomographic reconstruction from 2D US images presented and tested are:

- **Base-Hard**. This is the model previously described in Sec. 3.1. It is based on the architecture described in Sec. 3.2.1. The model has the same architecture as described by Prevost *et al.* in [32], and it is tested to understand if the implementation this architecture to reconstruct mobile and deformable anatomies (liver) is feasible.
- **Base-Soft**. This model is the one described in Sec. 3.1. This model was tested to understand which procedure to remove irrelevant data from the US images performed best.
- **MultiTask**. This model is based on the architecture presented in Fig. 3.10. The model was tested to understand if the integration of optical flow information in a different way could improve the reconstruction of the US volumes.
- **Attention-0.5**. This model is the one described in Sec. 3.3.1. The model is tested to understand how the addition of a TSGA map in the model input affect the reconstruction and if it is feasible for US reconstruction from liver intraoperative US images. In this case, the TSGA map added to its input is multiplied by a 0.5 ratio. Thus the value of a 255 in a pixel is divided by 2, resulting in a value of 127 (rounded down).
- **Attention-1.0**. This model is the one described in Sec. 3.3.1. For this model, the TSGA map is given its full weights, meaning that the map is multiplied by a ratio of 1.0. Thus if a pixel has value 255 in the computed map, it is given as input to the model as 255.

The models were trained on a machine with two Nvidia GTX 1080Ti GPUs, Intel Xeon CPU model E-2630 and 128 Gigabytes of RAM. A training of 100 epochs took around 68 hours to complete for the Attention-* models and 55 for the Base-Hard and MultiTask. The Base-Soft model took 74 hours to complete its 100 epochs training.

4.2.1 Hyper-parameter Optimization



(a) Base-Hard training loss. (b) Attention-0.5 training loss. (c) Attention-1.0 training loss.

Figure 4.3: Loss functions over number of epochs of the standardized training sessions of three of the presented models.

Figure 4.3 illustrates the training losses for the Base-Hard, Base-Soft, Attention-1.0 and Attention-0.5 models. From an analysis of the loss functions, it was not clear when the training was converging. Considering that the output of the model converges after a certain number of epochs (variable depending on the model trained), and that it does not appear from an analysis of the training loss when it happens, the following approach was followed to find the best training hyper-parameters.

At first all models were trained on the dataset described in Sec. 3.1 with the same parameters: a training of 100 epochs with an Adam optimizer set with a learning rate of $1 \cdot 10^{-4}$ and no decay. At the end of each epoch, a checkpoint of the trained model is stored. Then, starting from the training checkpoint at 100 epochs, the models were used to reconstruct a sweep from the test dataset. The checkpoints of the model at every 5 epochs (descending) were used to reconstruct the same sweep until the output differed from the previous step. If the number of epochs where the reconstructed volumes differed was smaller then 50, then the learning rate of the optimizer was scaled by 10^{-1} and the training restarted. If otherwise, a new training was started with decay set to 0.9. Once this last training finished, 10 volumes from the "Average" group (see. Sec. 3.1) in the test dataset were reconstructed and quantitatively analyzed with the same 8 metrics used for the experiment on patient data. The best performing checkpoint was chosen from the comparison of these results.

The models listed in the previous section were trained with the following values after hyper-parameter optimization.

Base-Hard: The model presented and studied in the following tests has been trained using MSE loss as function on the train set for a total of 35 epochs. The optimizer used for the training was Adam with learning rate set to 0.0001 that started decaying after 10 epochs by 0.9.

Base-Soft: The model presented and studied in the following tests, has been trained using MSE loss as function on the train set for a total of 35 epochs. The optimizer used for the training was Adam with learning rate set to 0.0001 that started decaying after 10 epochs by 0.9.

MultiTask: The Multi-task task model was trained for both tasks using MSE as loss function. The training was performed in an alternate fashion. The weights in the downscaling convolution were shared between the two tasks during the training. Each task has been trained for 5 rounds of 15 epochs for each task, resulting in 60 epochs of training for each task and 120 in total. The optimizer used was again Adam with different learning rates for the two tasks. For the training of both tasks, the learning rate was set to 0.0001 with a decay of 0.9 that started at the first epoch.

Attention-0.5: The Attention-0.5 model was trained using MSE as loss function for 90 epochs on the train set. The optimizer used for the training was Adam with learning rate set to 0.00001 with decay of 0.9 after 10 epochs.

Attention-1.0: The Attention-1.0 model was trained using MSE as loss function for a total of 60 epochs on the train set. The optimizer used for the training was Adam with learning rate set to 0.00001 with decay of 0.9 after 10 epochs.

All of the presented models were trained with MSE as loss function on the train dataset (see Sec. 3.1). The Mean Squared Error (MSE) was used as loss function since it obtained good performances in literature for US tomographic reconstruction ([32], [15]) and for optical flow estimation ([9], [29]).

4.3 Results

The following sections will present the results obtained from the two experiments described in Section 4.1 on the models trained as indicated in Section 4.2.1. First, the values obtained with the US phantom measure followed by those acquired with the experiment on patient data will be presented and analyzed. A qualitative analysis of the reconstructions obtained with models will conclude the section.

4.3.1 Results - US Phantom Measure

The results of the US phantom measure are presented in Table 4.1. All the values in the table are in millimeters.

	In-Plane				Out-of-Plane			
	Average	Absolute	1Q	3Q	Average	Absolute	1Q	3Q
EM	0.42	1.09	0.29	0.65	0.20	0.80	0.50	1.20
Base-Hard	0.40	2.16	0.13	0.56	5.82	15.10	2.35	9.73
Base-Soft	0.51	2.18	0.15	0.61	6.61	20.47	4.03	9.98
MultiTask	0.50	1.97	0.35	0.55	5.23	17.97	3.41	8.20
Attention-0.5	0.48	1.66	0.04	0.55	4.93	13.42	1.48	7.04
Attention-1.0	0.50	2.15	0.43	0.43	4.87	7.84	3.50	6.09

Table 4.1: Results of US phantom measure experiment of the proposed models and EM-tracking. All values are in mm.

From the comparison of the results in average and absolute error for in-plane and motion estimation that all the presented models performed similarly when reconstructing the vertical line of beacons. This means that the in-plane motion estimation between slices was predicted accurately by all the reconstruction models. All the proposed models scored similarly, with slightly worse results when compared to EM tracking. The most accurate from the proposed models are the Base-Hard and the Attention-0.5 models with average errors of respectively 0.40, 0.48 mm and absolute of 2.16 and 1.66 mm. These results were expected since it was accurately estimated by other methods proposed in the literature of this field.

In contrast with in-plane motion estimation, the error values reported in the table for out-of-plane motion estimation shows a wider distribution across the presented models. From the results reported in the table, it appears that the most accurate of the presented models at estimating out-of-plane motion are the two TSGA augmented models, Attention-0.5 and Attention-1.0. They scored respectively average errors of 4.93 and 4.87 mm, and with absolute errors of 13.42 and 7.84 mm. Attention-1.0 outperforms all the other presented models in out-of-plane motion estimation with the smallest absolute and average errors. Therefore, the results suggests that the addition of a TSGA map improves the out-of-plane motion estimation over the Base-* models.

Furthermore, from a comparison of the results of the two Attention-* models it appears that, although the average error score is similar in the two models, Attention-1.0 manages to achieve a

much lower absolute error. On the other hand, when it comes to in-plane motion estimation, the Attention-0.5 achieved smaller errors. This suggests that the addition of TSGA knowledge improves out-of-plane motion estimation, but that it might compromise the in-plane estimation.

From a comparison of the results achieved by Base-Hard and Base-Soft, it appears that Base-Hard achieves smaller errors across all the metrics. As a result, from this experiment it appears that the use of hard attention was the better choice to remove irrelevant information from the US images, hence in the adaptation process. From a comparison of the results achieved by Base-Hard and MultiTask, the table shows that training the base model in a multi-task fashion slightly improved the out-of-plane estimation. Also to be noted that, on the contrary of the in-plane estimation, the comparison between the results achieved with the presented models and the EM tracking based reconstruction shows a significant difference in performances, seeing the EM tracking based reconstruction as the most accurate.

4.3.2 Results - Experiment on Patient Data

The results of the qualitative analysis of the methods on patient data are summarized in Table 4.2. The values shown in the table are all in millimeters aside for the three rotation values that are in degrees.

	PrevBase	Base-Hard	Base-Soft	MultiTask	Attention-0.5	Attention-1.0
t_x	3.54	7.71±0.23	8.02±0.27	6.83±0.23	3.98±0.21	4.40±0.13
t_y	3.05	5.17±0.11	6.10±0.17	4.44±0.11	3.92±0.16	4.05±0.11
t_z	4.19	5.33±0.20	6.96±0.22	5.06±0.21	4.73±0.20	4.24±0.20
r_x	2.63	3.59±0.12	3.66±0.12	3.20±0.11	2.57±0.10	2.47±0.12
r_y	2.52	3.14±0.10	3.97±0.16	2.89±0.10	2.93±0.10	2.59±0.12
r_z	1.93	3.92±0.13	4.00±0.18	2.74±0.14	3.43±0.15	2.45±0.12
Final Drift	14.40	79.70±26.13	86.33±31.54	72.58±25.63	51.86±15.12	56.11±19.90
Length	n/a	36.22±7.15	41.59±9.95	51.86±6.36	28.82±4.21	24.61±3.99

Table 4.2: Results of the experiment on patient data and those reported by Prevost *et al.* (e.g., PrevBase). The values are all in mm aside of the $r_{x,y,z}$ values that are in degrees.

For PrevBase, only 7 out of the 8 metrics are present in the table since Prevost *et al.* in [32] did not report the length error for their CNN with optical flow.

From the results in the table, a first comparison between the Base-Hard and Base-Soft models can be done: Base-Hard outperforms Base-Soft across all error metrics, proving once again that the implementation of hard attention to remove irrelevant information from the convex US slices was the better choice. Since the Base-* models were the first two to be implemented and tested, these results were known before the implementation and training of the other presented models. Consequently, the other presented models were only trained with the use of hard attention to remove irrelevant information from the convex US images.

An analysis of the average absolute errors relative to in-plane translation estimation (e.g. t_x and t_y) reported in the table, it appears that Attention-0.5 achieved the lowest values with respectively

3.98±0.21 mm and 3.92±0.16 mm. Attention-0.5 also resulted to be the best performing in terms of drift (e.g. propagation of in-plane estimation error), suggesting that it was overall the best performing of the presented models at estimating in-plane motion.

The values reported in the table for average absolute error of out-of-plane translation estimation (e.g. t_z) suggests that the best performing among the presented models is Attention-1.0 with the smallest value of 4.24±0.20 mm. The same outcome is also reported for the length error (e.g. the error propagation of out-of-plane translation estimation) values: Attention-1.0 is the best performing among the presented models.

An analysis of the results of the Attention-* models from both experiments suggests that there is a trend in the in-plane and out-of-plane estimation performances between these two models.

From a comparison between the results reported in the table for Base-Hard and MultiTask, it appears that the implementation of multi-task learning slightly improves the reconstruction performances of the Base-* models, scoring smaller errors across all metrics.

The comparison of the average absolute errors achieved by the presented models and those reported by Prevost *et al.* in [32] shows that the Attention-* models scored results that are almost on-par with PrevBase. However, when it comes drift and length errors, the presented models achieves much worse results: Attention-0.5 scores 51.86 ±15.12 mm for drift when Prevost *et al.* reports an error of only 14.40 mm.

4.3.3 Qualitative Analysis

The following Figures (4.4, 4.5), illustrate two tomographic reconstructions using Attention-0.5 as reconstruction method compared to their reconstruction performed with the EM tracking data (ground truth). Both deep-learning based reconstructions are generated with the Attention-0.5 model since it was one of the best performing in the experiment on patient data. Figure 4.4 illustrates the reconstruction of a volume from the "Good" category in the test set. This reconstruction can be considered as a best-case tomographic reconstruction. Reconstructions like this happened on ∼20% (all from the "Good" and "Average" categories) of the test dataset. The reconstruction of this sweep performed with EM tracking data was one of the most straight and less bumpy acquisition present in the dataset. An analysis of the slices in this figure indicates that the relevant anatomies (hepatic system and a tumor visible in both vertical and horizontal slices) are accurately reconstructed. Also, not many artifacts are present in the US volume reconstructed with Attention-0.5.

Figure 4.5 shows a worst-case reconstruction of an "Average" volume from the test set (see Sec. 3.1). Situations like the one illustrated in the Figure happened for ∼25% (all from the "Average" category but one from the "Good") of the reconstructed volumes in the test set.

The resulting reconstructed volume is completely different from its ground truth. Although the in-plane motion estimation is still comparable to the ground truth, the out-of-plane motion estimation is totally inaccurate. The volume reconstructed with EM indicates that the sweep was fairly linear with some bumps in the mid portion of the volume. However, the Attention-0.5 based reconstruction not only presents an upward trend in the vertical motion estimation (which is not

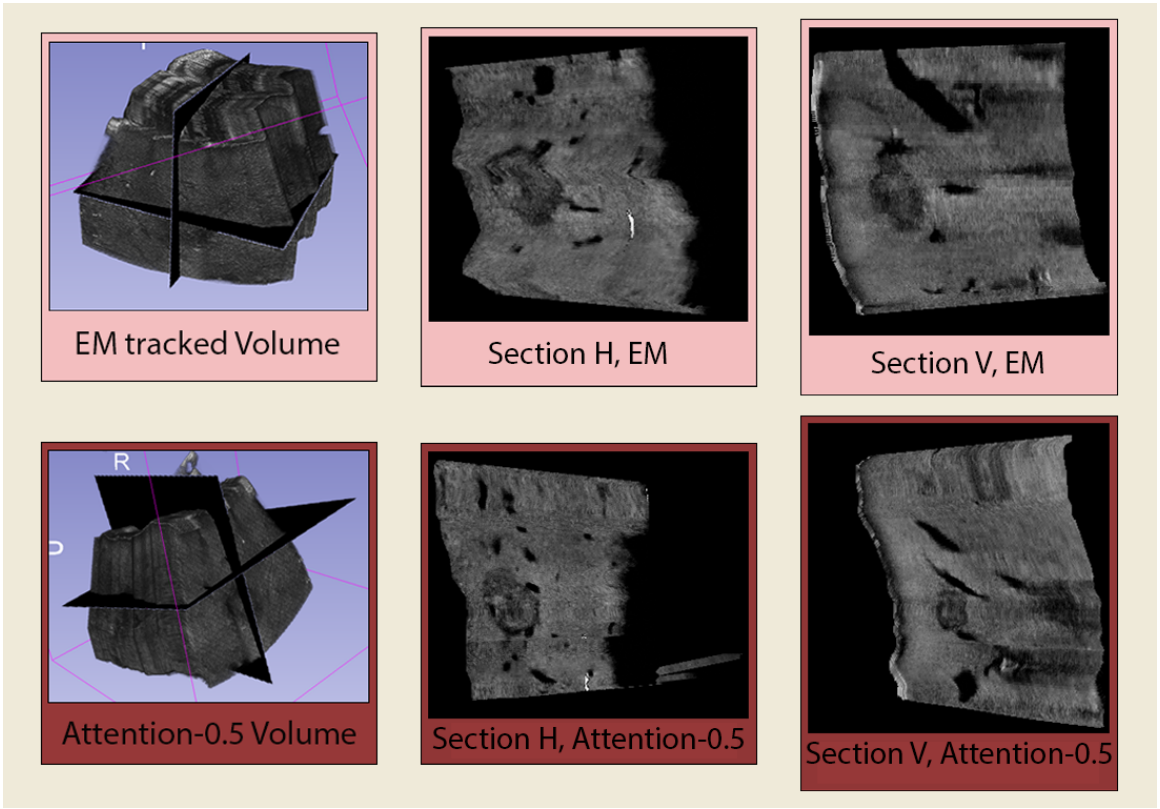


Figure 4.4: Example of best-case reconstruction of a "Good" volume from the test set. The three images on top represents the reconstructed volume with horizontal and vertical section for the EM based reconstruction. The three images at the bottom are the same for the volume reconstructed with Attention-0.5.

present in the EM based reconstruction), but also the out-of-plane estimation is negative in the mid portion of the volume. A diagonal line can be seen on the left side of the horizontal slice of the Attention-0.5 reconstruction illustrated in Fig. 4.5. The images in that portion of the sweep were positioned by Attention-0.5 on top of the previous ones, which means that a negative motion was estimated between US images. As a result, the volume reconstructed with Attention-0.5 presents a portion where the images overlap but should not. Also, as it is shown in the section planes, the scanned anatomies are not accurately reconstructed and the structure of the scanned ROI is not clear.

4.4 Limitations and Future Work

During the evaluation with the two experiments described in Section 4.1 of the models presented in Chapter 3 some unexpected behaviors appeared.

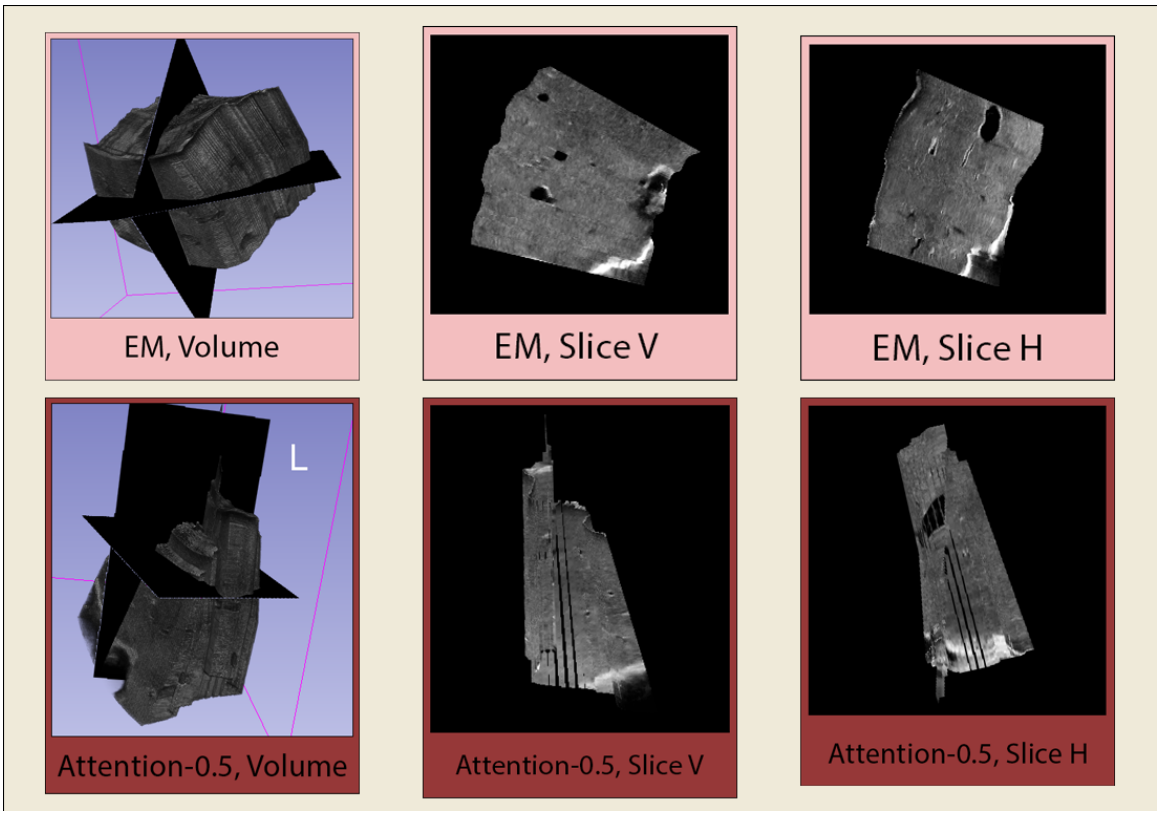
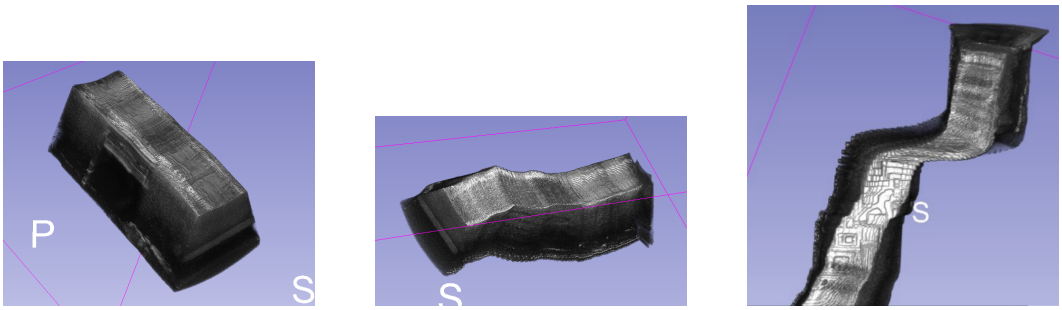


Figure 4.5: Example of worst-case reconstruction of an "Average" volume from the test set. The three images on top represents the reconstructed volume with horizontal and vertical section for the EM based reconstruction. The three images at the bottom are the same for the volume reconstructed with Attention-0.5.

Figure 4.6 illustrates a situation where the proposed models failed at following the trajectory of the US phantom sweep. There, three tomographic reconstructions obtained with the EM tracking data and the Attention-* models are pictured. It can be seen how in both the reconstructions achieved with the Attention-* models, the volumes diverged from the ground truth approximately in the same point (images 195-220 out of 650 in the sweep). In specific, the volume reconstructed with the Attention-1.0 diverges from the ground truth by 64 mm and the one obtained with Attention-0.5 by 52 mm. After a visual inspection of the US images in the portion of the US phantom sweep where the two reconstructions diverged from the ground truth trajectory, it appeared that the portion of the phantom was the one in Figure 4.7b where the Us images resembled the one in Figure 4.7a.

Therefore, we believe that the reconstruction problem illustrated in Figure 4.6 and just discussed is related to the addition of transducer-specific geometry knowledge. Not only this situation was not present in the reconstructions obtained with the other models, but also the Attention-1.0 based reconstruction was more affected than the Attention-0.5 one. The US image illustrated in Fig. 4.7a has almost no relevant information in the middle, and considering that the these are cropped



(a) US phantom sweep reconstruction with EM tracking data.

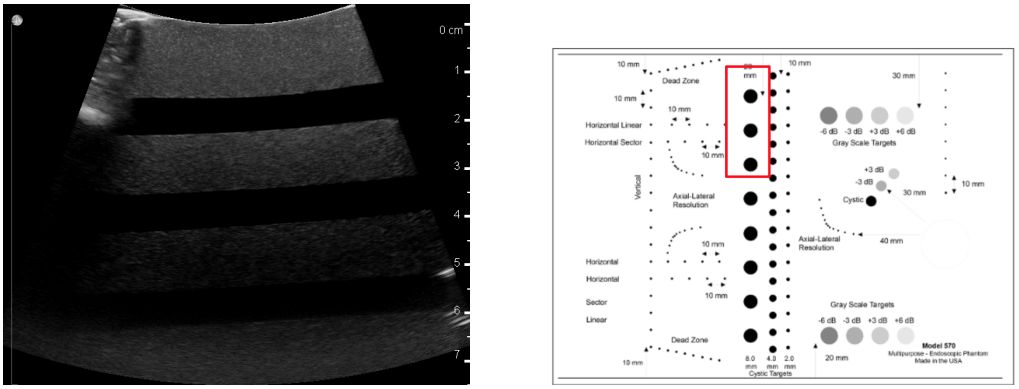
(b) US phantom sweep reconstruction with Attention-0.5.

(c) US phantom sweep reconstruction with Attention-1.0.

Figure 4.6: Reconstructed US volumes of the sweep performed on the ATS US phantom model 539.

from the center and that the additional geometry knowledge focuses the network on the motion in the central part of the input, the input was lacking critical information where the network was paying more attention. Furthermore, the shape of the scanned inserts of the phantom and the speckle around them were following a side-wise motion. As a result, the reconstruction followed the incorrect trajectory estimated from the synthetic anatomy in the phantom and accentuated by the attention map. It has to be considered that situations like the one illustrated by Fig. 4.7a are not common in patient data. Consequently, we believe that this does not affect as much the models when reconstructing US volumes from patient data.

Although Attention-1.0 was the best performing model at estimating out-of-plane displacement in both experiments, Attention-0.5 was the model that had best scores for in-plane motion estimation in the experiment on patient data. These two results suggests the existence of a trend between the



(a) Example US image in the portion of the phantom highlighted in Fig. 4.7b.

(b) Highlighted the portion of the US phantom.

Figure 4.7: The US images and portion of the US phantom where the Attention-* models failed their displacement estimation.

weight of the TSGA map and the ability of in- and out-of-plane motion estimation. In specific, since Attention-0.5 was better at estimating in-plane motion and Attention-1.0 out-of-plane and that the only difference between the two models are the weights given to the TSGA map, it appears that the aforementioned trend is caused by the intensity of the TSGA map. Consequently, we plan in the future to train models where the TSGA map is integrated with weights multiplied by a ratio that ranges between 0.5 and 1.0 (e.g., 0.6, 0.7, ...) to better assess this trend.

For future work, we plan to investigate other ways to integrate TSGA in the models. It was demonstrated that incorporating TSGA to the model as a map concatenated to the input improves the tomographic reconstruction performance, however there are other ways that can be adopted to do so. We believe that, instead of concatenating the TSGA map to the input image, scaling the optical flow information (e.g., multiplying the optical flow by a 0 to 1 TSGA map and use that as input) could improve the results. In the Attention-* models the TSGA knowledge was incorporated as additional channel. Given that Base-Hard outperformed Base-Soft and Base-Hard had less channels in the input image, we believe that avoiding the additional channel required to integrate the TSGA map could improve the accuracy of the tomographic reconstructions.

The presented models were trained with the use of an Adam optimizer which is renown to be fast at converging the loss function but not the best at finding its minimum. Furthermore, the values of the training loss were fluctuating across the epochs, not following a steady decrease. As a consequence, the implementation of a different optimizer could help to narrow the distribution of the quality of the reconstructed volumes (see Fig. 4.4 and 4.5). In the specific case, we plan to perform a training of the presented models with the Adagrad optimizer, since it was reported by Prevost *et al.* that the model trained with it was achieving the best results.

Figure 4.8 illustrates an incorrect behaviour of the proposed tomographic reconstruction methods. From the visual inspection of the reconstructed volumes in the test set, it appeared that in the near

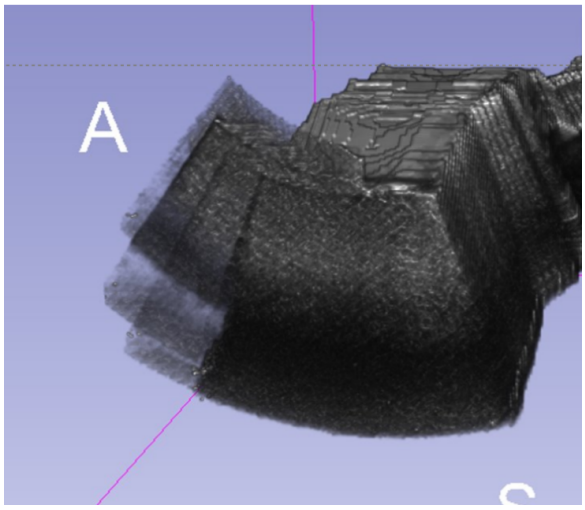


Figure 4.8: Detail of a US sweep of a patient reconstructed with Attention-0.5.

totality of the cases (30 out 34 volumes qualitatively evaluated), the images at the beginning of the sweeps were wrongly placed by the proposed models as depicted in the Figure. We believe that this issue is generated by the loss function used for training of the presented models (e.g., MSE), since the same behaviour was previously reported in other studies on methods for tomographic reconstruction from 2D US data ([47], [19]). Guo *et al.* [15] proposed a method for tomographic reconstruction of 2D US images, where they also presented and implemented what the authors defined Case-Wise Correlation Loss. They stated that models trained with this loss function should not suffer from the problem illustrated in Fig. 4.8. Therefore, for future work, we plan to implement a model that can make use of such loss function to test if the aforementioned problem can be avoided.

From the results of the two experiment analyzed in the previous sections, it was not clear if the implementation of multi-task learning for both optical flow and pair-wise displacement estimation improved the Base-Hard model. Therefore, given that in this study multi-task learning was only assessed with alternate training, it is suggested to do further research implementing a joint multi-task training. Joint multi-task learning is a different approach on the minimization of the multiple loss functions of the tasks. The model is trained with an additional optimizer which is responsible to minimize the loss functions of the two tasks, instead of alternatively train it between them. Given that this technique was adopted in the training of ActionFlowNet ([28]), we believe that it could improve the MultiTask model reconstruction abilities.

Given that both the addition of TSGA knowledge in the form of a map and the implementation of multi-task learning (even if multi-tasking requires a more further investigation) resulted to improve the reconstruction abilities of the Base-Hard model, for future work it is suggested to implement a model that implmenets multi-task learning and integrates TSGA.

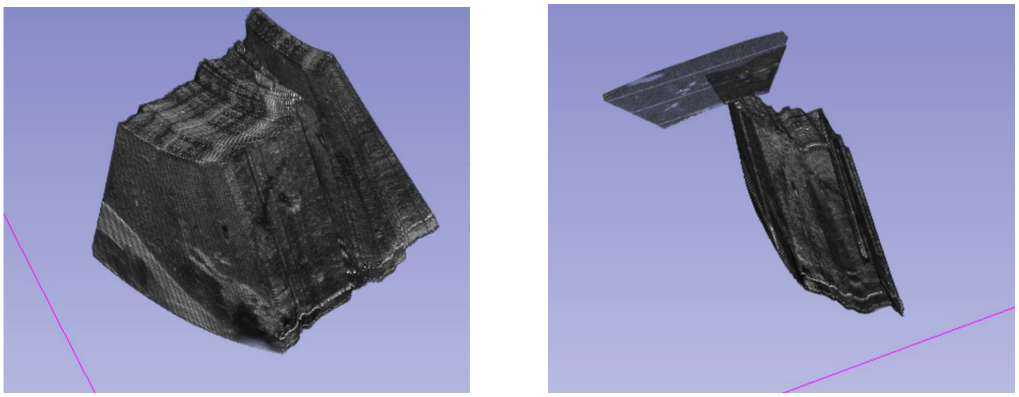
Since the estimations of the presented models are based on speckle motion, these are only able to correctly reconstruct from a sweep where rotations are small and out-of-plane motion along depth is one-directional [32]. As a result, these models are not able of accurately reconstruct volumes from sweeps where the same area is scanned repetitively, or if the transducer is rotated around the vertical axis on the same area. No solution to this problem has been currently reported, to our knowledge, in the literature of this field. A solution could be to base the reconstruction on the anatomies present in the 2D US images, which is not taken into consideration by the presented models (or at least it is not their primary focus since it was reported that the anatomy in the US images used for the training still influences the generalization of the reconstruction method [32]). Therefore, a model that bases its estimation on the anatomy present in the US images thanks to an anatomy-driven attention map could be implemented. This attention map could be extracted from from the segmentation of relevant anatomies in the 2D US images. Then, this map could be integrated into a model to make it learn the generalized anatomical shape of the liver. Such model would be able to estimate the position and rotation of a single US image analyzing the anatomies illustrated in that image. This process would resemble the analysis that radiologists perform when assessing 2D US images. They are trained to orient themselves in 3D space from a single 2D image, and their assessment is based (along with other factors) on the anatomy present in the analyzed US image.

4.5 Conclusion

With the results of the evaluation process of the presented models, it is now possible to answer the research question and sub-questions defined in 1.1.

In the previous sections, the adaptation of the model proposed by Prevost *et al.* on the dataset described in Sec. 3.1 and two variations of it were assessed. With a direct comparison of the results of the presented Base-Hard model and PrevBase it was shown that Base-Hard can reconstruct US volumes from 2D US images with worse accuracy. This was expected due to the additional challenges presented in Sec 3.2.3. The deformability and complex anatomy of the liver, the smaller dataset and the different US transducer all add to the already difficult challenge of tomographic reconstruction from US data. Nontheless, the Base-Hard model demonstrated to be able to reconstruct volumes from 2D US images of the liver as it is illustrated in Fig. 4.9. Therefore, the first research sub-Question 1 can be positively answered: it is feasible to adapt the work from Prevost *et al.* to reconstruct mobile and deformable anatomies.

The results of the two experiments also indicated that incorporating the optical flow information to the network in a multi-task fashion can slightly improve the performances of the Base-Hard model. Although it appeared a positive trend in the displacement estimation abilities of MultiTask over Base-Hard, the reported results are not statistically relevant to give a final answer to 3. Consequently, further investigation is required to give a final answer to sub-Question 3.



(a) Reconstruction with EM data. (b) Reconstruction with Base-Hard.

Figure 4.9: The reconstruction of the sweep on a patient.

From the comparison of the results achieved by Base-Hard and the Attention-* models, it appeared that augmenting the input of the Base-Hard model with transducer-specific geometry map improved the accuracy of the tomographic reconstructions. Both the geometry knowledge augmented models, Attention-0.5 and Attention-1.0, outperformed Base-Hard, achieving results for pair-wise absolute average errors almost on-par with those presented by Prevost *et al.*. Thus, sub-Question 2 is positively answered: it is feasible to improve the Base-Hard model with the integration of transducer-specific geometry attention.

The models for tomographic reconstruction of 2D US images of the liver presented in this study demonstrated that are able to reconstruct volumes. Their accuracy is however still far from being clinically accepted (e.g., errors are greater then 5mm as stated in Sec. 1.1), making them not yet suited for liver surgery. As a result, the main research question of this study can be answered as follows: tomographic reconstruction from 2D intraoperative US images of the liver is feasible but further research is required to achieve a clinically accepted accuracy of the reconstructions.

Bibliography

[1] Askeland, C., Solberg, O. V., Bakeng, J. B. L., Reinertsen, I., Tangen, G. A., Hofstad, E. F., Iversen, D. H., Våpenstad, C., Selbekk, T., Langø, T., Hernes, T. A. N., Leira, H. O., Unsgård, G., and Lindseth, F. CustusX: an open-source research platform for image-guided therapy. *International Journal of Computer Assisted Radiology and Surgery 11*, 4 (Sept. 2015), 505–519.

[2] Bax, J., Cool, D., Gardi, L., Knight, K., Smith, D., Montreuil, J., Sherebrin, S., Romagnoli, C., and Fenster, A. Mechanically assisted 3d ultrasound guided prostate biopsy system. *Medical Physics 35*, 12, 5397–5410.

[3] Buades, A., Coll, B., and Morel, J. M. A review of image denoising algorithms, with a new one. *Multiscale Modeling & Simulation 4*, 2 (Jan. 2005), 490–530.

[4] Chen, J.-F., Fowlkes, J. B., Carson, P. L., and Rubin, J. M. Determination of scan-plane motion using speckle decorrelation: Theoretical considerations and initial test. *International Journal of Imaging Systems and Technology 8*, 1 (1997), 38–44.

[5] Chen, X., Wen, T., Li, X., Qin, W., Lan, D., Pan, W., and Gu, J. Reconstruction of freehand 3d ultrasound based on kernel regression. *BioMedical Engineering OnLine 13*, 1 (2014), 124.

[6] Chen, Z., and Huang, Q. Real-time freehand 3d ultrasound imaging. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization 6*, 1 (May 2016), 74–83.

[7] CIRS. Ats model 539, us phantom. `https://www.cirsinc.com/products/ultrasound/ats-urethane/multi-purpose-phantom/`.

[8] de Lambert, A., Esneault, S., Lucas, A., Haigron, P., Cinquin, P., and Magne, J.-L. Electromagnetic tracking for registration and navigation in endovascular aneurysm repair: A phantom study. *European Journal of Vascular and Endovascular Surgery 43*, 6 (June 2012), 684–689.

[9] Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., van der Smagt, P., Cremers, D., and Brox, T. Flownet: Learning optical flow with

convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (December 2015).

[10] Duda, R., and Hart, P. *Pattern classification and scene analysis.* 1973.

[11] Fenster, A., Parraga, G., and Bax, J. Three-dimensional ultrasound scanning. *Interface Focus 1*, 4 (2011), 503–519.

[12] Gao, F., Yoon, H., Wu, T., and Chu, X. A feature transfer enabled multi-task deep learning model on medical imaging. *Expert Systems with Applications 143* (2020), 112957.

[13] Gobbi, D. G., and Peters, T. M. Interactive intra-operative 3d ultrasound reconstruction and visualization. In *Medical Image Computing and Computer-Assisted Intervention — MICCAI 2002* (Berlin, Heidelberg, 2002), T. Dohi and R. Kikinis, Eds., Springer Berlin Heidelberg, pp. 156–163.

[14] Gonzalez, R. C., and Woods, R. E. *Digital Image Processing (3rd Edition).* Prentice-Hall, Inc., USA, 2006.

[15] Guo, H., Xu, S., Wood, B., and Yan, P. Sensorless freehand 3d ultrasound reconstruction via deep contextual learning. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020* (Cham, 2020), A. L. Martel, P. Abolmaesumi, D. Stoyanov, D. Mateus, M. A. Zuluaga, S. K. Zhou, D. Racoceanu, and L. Joskowicz, Eds., Springer International Publishing, pp. 463–472.

[16] Huang, Q. Speckle suppression and contrast enhancement in reconstruction of freehand 3d ultrasound images using an adaptive distance-weighted method. *Applied Acoustics 70*, 1 (2009), 21 – 30.

[17] Huang, Q., and Zeng, Z. A review on real-time 3d ultrasound imaging technology. *BioMed Research International 2017* (2017), 1–20.

[18] Huang, Q., Zheng, Y., Lu, M., Wang, T., and Chen, S. A new adaptive interpolation algorithm for 3d ultrasound imaging with speckle reduction and edge preservation. *Computerized Medical Imaging and Graphics 33*, 2 (2009), 100 – 110.

[19] Johnson, J., Alahi, A., and Fei-Fei, L. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision – ECCV 2016* (Cham, 2016), B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., Springer International Publishing, pp. 694–711.

[20] Lindseth, F., Langø, T., Selbekk, T., Hansen, R., Reinertsen, I., Askeland, C., Solheim, O., Unsgård, G., Mårvik, R., and Hernes, T. *Ultrasound-Based Guidance and Therapy.* 06 2013.

[21] Lu, X., Zhang, S., Yang, W., and Chen, Y. SIFT and shape information incorporated into fluid model for non-rigid registration of ultrasound images. *Computer Methods and Programs in Biomedicine 100*, 2 (Nov. 2010), 123–131.

[22] McCann, H., Sharp, J., Kinter, T., McEwan, C., Barillot, C., and Greenleaf, J. Multidimensional ultrasonic imaging for cardiology. *Proceedings of the IEEE 76*, 9 (1988), 1063–1073.

[23] Medical, B. Us intraoperative t-shaped transducer model i14c5t. `https://www.bkmedical.com/transducers/i14c5t-t-shaped-intraoperative/`.

[24] Moeskops, P., Wolterink, J. M., van der Velden, B. H. M., Gilhuijs, K. G. A., Leiner, T., Viergever, M. A., and Išgum, I. Deep learning for multi-task medical image segmentation in multiple modalities. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016* (Cham, 2016), S. Ourselin, L. Joskowicz, M. R. Sabuncu, G. Unal, and W. Wells, Eds., Springer International Publishing, pp. 478–486.

[25] Mohamed, F., Mong, W., and Yusoff, Y. Quaternion based freehand 3d baby phantom reconstruction using 2d ultrasound probe and game controller motion and positioning sensors. pp. 272–278.

[26] Mohamed, F., and Siang, C. V. A survey on 3d ultrasound reconstruction techniques. In *Artificial Intelligence - Applications in Medicine and Biology*. IntechOpen, July 2019.

[27] NDI. Aurora em tracking system for freehand us. `https://www.ndigital.com/products/aurora/`.

[28] Ng, J. Y., Choi, J., Neumann, J., and Davis, L. S. Actionflownet: Learning motion representation for action recognition. *CoRR abs/1612.03052* (2016).

[29] Ng, J. Y.-H., Choi, J., Neumann, J., and Davis, L. S. ActionFlowNet: Learning motion representation for action recognition. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)* (Mar. 2018), IEEE.

[30] Øye, O. K., Wein, W., Ulvang, D. M., Matre, K., and Viola, I. Real time image-based tracking of 4d ultrasound data. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2012* (Berlin, Heidelberg, 2012), N. Ayache, H. Delingette, P. Golland, and K. Mori, Eds., Springer Berlin Heidelberg, pp. 447–454.

[31] Pezzotti, N., de Weerdt, E., Yousefi, S., Elmahdy, M. S., van Gemert, J., Schülke, C., Doneva, M., Nielsen, T., Kastryulin, S., Lelieveldt, B. P. F., van Osch, M. J. P., and Staring, M. Adaptive-cs-net: Fastmri with adaptive intelligence, 2019.

[32] Prevost, R., Salehi, M., Jagoda, S., Kumar, N., Sprung, J., Ladikos, A., Bauer, R., Zettinig, O., and Wein, W. 3d freehand ultrasound without external tracking using deep learning. *Medical Image Analysis 48* (2018), 187 – 202.

[33] Raza, A. Hepatocellular carcinoma review: Current treatment, and evidence-based medicine. *World Journal of Gastroenterology 20*, 15 (2014), 4115.

[34] Schers, J., Troccaz, J., Daanen, V., Fouard, C., Plaskos, C., and Kilian, P. P6d-1 3d/4d ultrasound registration of bone. In *2007 IEEE Ultrasonics Symposium Proceedings* (2007), pp. 2519–2522.

[35] Schneider, R. J., Perrin, D. P., Vasilyev, N. V., Marx, G. R., del Nido, P. J., and Howe, R. D. Real-time image-based rigid registration of three-dimensional ultrasound. *Medical Image Analysis 16*, 2 (Feb. 2012), 402–414.

[36] Solberg, O. V., Langø, T., Tangen, G. A., Mårvik, R., Ystgaard, B., Rethy, A., and Hernes, T. A. N. Navigated ultrasound in laparoscopic surgery. *Minimally Invasive Therapy & Allied Technologies 18*, 1 (Jan. 2009), 36–53.

[37] Solberg, O. V., Lindseth, F., Bø, L. E., Muller, S., Bakeng, J. B. L., Tangen, G. A., and Hernes, T. A. N. 3d ultrasound reconstruction algorithms from analog and digital data. *Ultrasonics 51*, 4 (2011), 405–419.

[38] Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., and Bray, F. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians 71*, 3 (Feb. 2021), 209–249.

[39] Tao, Z., Tagare, H. D., and Beaty, J. D. Evaluation of four probability distribution models for speckle in clinical cardiac ultrasound images. *IEEE Transactions on Medical Imaging 25*, 11 (2006), 1483–1491.

[40] Thomson, B. Automated vascular region segmentation in ultrasound to utilize surgical navigation in liver surgery. `http://essay.utwente.nl/79399/`, August 2019.

[41] Tsim, N. C. Surgical treatment for liver cancer. *World Journal of Gastroenterology 16*, 8 (2010), 927.

[42] Viola, P., and Wells, W. M. Alignment by maximization of mutual information. In *Proceedings of IEEE International Conference on Computer Vision* (1995), pp. 16–23.

[43] Wang, Z., Slabaugh, G., Unal, G., and Fang, T. Registration of ultrasound images using an information-theoretic feature detector. In *2007 4th IEEE International Symposium on Biomedical Imaging: From Nano to Macro* (2007), pp. 736–739.

[44] Wen, T., Li, L., Zhu, Q., Qin, W., Gu, J., Yang, F., and Xie, Y. Gpu-accelerated kernel regression reconstruction for freehand 3d ultrasound imaging. *Ultrasonic Imaging 39*, 4 (2017), 240–259. PMID: 28627330.

[45] Wen, T., Yang, F., Gu, J., and Wang, L. A novel bayesian-based nonlocal reconstruction method for freehand 3d ultrasound imaging. *Neurocomputing 168* (2015), 104 – 118.

[46] Wen, T., Zhu, Q., Qin, W., Li, L., Yang, F., Xie, Y., and Gu, J. An accurate and effective fmm-based approach for freehand 3d ultrasound reconstruction. *Biomedical Signal Processing and Control 8*, 6 (2013), 645 – 656.

[47] Yang, Q., Yan, P., Zhang, Y., Yu, H., Shi, Y., Mou, X., Kalra, M. K., Zhang, Y., Sun, L., and Wang, G. Low-dose CT image denoising using a generative adversarial network with wasserstein distance and perceptual loss. *IEEE Transactions on Medical Imaging 37*, 6 (June 2018), 1348–1357.

[48] Yen, J. T., Steinberg, J. P., and Smith, S. W. Sparse 2-d array design for real time rectilinear volumetric imaging. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control 47*, 1 (2000), 93–110.