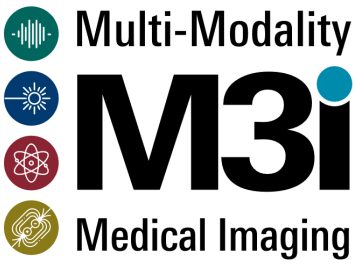
Deep learning segmentation of 3D ultrasound thyroid scans

A comparison of training strategies and 3D ultrasound acquisition methods

Roxane Munsterman, BSc



A thesis presented for the degree of Master of Science
Faculty of Science and Technology
Biomedical Engineering
University of Twente
06-2023

Examination committee:
Chair M3i: Prof. Dr. S. Manohar
Daily supervisor M3i: T. Boers, MSc
External member MAI: Dr. J.M. Wolterink

1 Abbreviations

- RLN Recurrent laryngeal nerve
- RFA Radiofrequency ablation
- JV Jugular vein
- CA Carotid artery
- GUI Graphical user interface
- FOV Field of view
- SRAD Speckle reducing anisotropic diffusion
- TIRADS Thyroid Imaging Reporting and Data System
- LA Laser ablation
- CNN Convolutional neural network
- FCN Fully convolutional neural network
- MONAI Medical Open Network for artificial intelligence
- HD_{95} Hausdorff distance 95^{th} percentile
- DSC Dice similarity coefficient
- mAP Mean average precision

2 Preface

My master thesis 'Deep learning segmentation of 3D ultrasound thyroid scans' is my final assignment to complete the master's of Biomedical Engineering. After my master's subjects and internship in the track Imaging and In Vitro Diagnostics, I came to the conclusion that my interests lie in the world of computer vision and image processing in medical imaging. With a broad research scope in many imaging modalities, I was sure the Multi-Modality Medical Imaging (M3i) group would have a project I would like to work on. My interest was caught with 3D ultrasound and after a discussion on possible master projects, I found my project. What I liked about the project was being able to work on the whole process, from making my own scans to how my research could be implemented in the clinic.

Through the entire project, I had great help and support from my daily supervisor, Tim Boers, who greatly helped shape the project and guide me through the process. I really appreciated your insights during our weekly meetings and your help in making decisions and formulating everything on paper. I would also like to thank my supervisor Srirang Manohar for his ideas and enthusiasm about the project and, together with the rest of the staff, making me feel very welcome at M3i.

The project meant taking a leap into the deep of deep learning. I would like to thank Jelmer Wolterink for his great teaching and for hosting the imaging colloquia and imaging meetings. I also got some great suggestions during these meetings, which helped me think of solutions to problems I ran into. A special thanks to Diewertje Alblas, Elina Thibeua-Sutre and Bruno De Santi for making some extra time to help me with my code problems.

I would also like to thank Sicco Braak for showing me the around at the radiology department and helping in gathering my dataset. I am also thankful for all the enthusiasm of the staff in helping me gathering the dataset.

And a special thanks to all the students in TL1332 who made working even more fun with all the nice coffee breaks and conversations. And a big thank you to my friends, family and my boyfriend for their amazing moral support. Finally, thank you to everyone who came by to let me scan your thyroids.

Abstract

Thyroid nodules have a high prevalence and can be detected upon ultrasound imaging. To improve the diagnosis and treatment of thyroid nodules, a 3D segmentation method for ultrasound scans was developed, segmenting the thyroid, carotid artery (CA), and jugular vein (JV). The goal of the method is to aid needle-based interventions, such as radiofrequency ablation (RFA) and improve volumetry accuracy. A tracked sweep dataset from an online repository was used together with a dataset acquired with matrix transducer, which allows for fast 3D volume acquisition. Both datasets consisted of ultrasound scans and annotations from 27 subjects. Pre-processing techniques were applied to enhance the scans, including voxel size normalization and speckle reduction. A U-Net was trained with different strategies (2D, 2.5D majority vote, and 3D) on both the matrix dataset and tracked sweep dataset, to find the best training strategy. The Dice similarity coefficient (DSC) and Hausdorff Distance 95% (HD_{95}) were used to assess the model's performance. The volume of the prediction was compared to the ground truth and to volumes obtained using the ellipsoid formula. The results showed variations in performance among the training strategies. The 2D model achieved the best results for the tracked sweep dataset in terms of median DSC (0.934, 0.924, 0.897) and HD_{95} (1.206, 0.588, 1.571 mm) for the thyroid, CA and JV respectively. For the matrix dataset, the 3D train strategy gave overall best results in its median DSC (0.869, 0.930, 0.856) and HD_{95} (1.814, 0.606, 1.405 mm) for the thyroid, CA and JV respectively. The model demonstrated lower median volume errors (4.45%) compared to the ellipsoid formula (13.84%) for thyroid volume estimation in the tracked sweep dataset. For the matrix dataset, an error of 7.40% was achieved. A graphical user interface was developed for visualization and clinical use of the segmentation results. A 3D segmentation method for ultrasound volumes of the thyroid, CA and JV was developed. This work paves the way for the development of a planning and navigation method to be used with RFA for thyroid nodules.

Reader

This thesis starts with an introduction chapter, briefly stating the current medical status of thyroid nodule pathology, diagnosis and treatment. Then the focus will be on segmentation and the role deep learning has played in the segmentation of thyroids so far. The introduction will end with the research goal and questions. Chapter 4 contains background information to learn more about the clinical background, the ultrasound imaging modality, or segmentation methods, ending with a deep learning section which will contain an explanation of all technical terms mentioned in this thesis. Chapter 5 contains the methods, followed by the results in Chapter 6 used to find an answer to the research goal and questions. Chapter 7 contains the discussion, discussing results, the limitations of the study, the clinical value of the results, and suggestions for future research. Chapter 8 summarises the results, leading to a conclusion.

Samenvatting

Schildklier nodi komen veel voor en kunnen worden gedetecteerd met ultrasound. Om de diagnose en behandeling van schildklier nodi te verbeteren, werd een 3D-segmentatiemethode voor ultrasound scans ontwikkeld die de schildklier, carotide (CA) en jugularis (JV) segmenteert. Het doel van de methode is om interventies met naalden, zoals radiofrequente ablatie (RFA), te ondersteunen en de nauwkeurigheid van volumetrie te verbeteren. Een publiek beschikbare tracked sweep dataset werd gebruikt samen met een dataset die was verkregen met een matrix transducer, die een snelle 3D volume acquisitie mogelijk maakt. Beide datasets bestonden uit ultrasound scans en annotaties van 27 proefpersonen. Er werden voorbewerkingstechnieken toegepast om de scans te verbeteren, waaronder voxelformaatnormalisatie en speckle reductie. Er werd een U-Net getraind met verschillende strategieën (2D, 2,5D majority vote en 3D) op zowel de matrixdataset als de tracked sweep dataset om de beste trainingsstrategie te vinden. De Dice similariteitscoëfficiënt (DSC) en Hausdorff Distance 95% (HD_{95}) werden gebruikt om de prestaties van het model te beoordelen. Het volume van de voorspelling werd vergeleken met de annotaties en met volumes verkregen met de ellipsoïde formule. De resultaten toonden variaties in prestaties tussen de trainingsstrategieën. Het 2D-model behaalde de beste resultaten voor de tracked sweep dataset in termen van mediaan DSC (0.934, 0.924, 0.897) en HD_{95} (1.206, 0.588, 1.571 mm) voor respectievelijk de schildklier, CA en JV. De 3D train strategie gaf de beste results voor de matrix dataset in mediaan DSC (0.869, 0.930, 0.856) en HD_{95} (1.814, 0.606, 1.405 mm) voor respectievelijk de schildklier, CA en JV. Het model toonde ook lagere mediane volume errors (4.45%) vergeleken met de ellipsoïde formule (13.84%) voor het berekenen van het schildkliervolume in de tracked sweep dataset. Voor de matrix dataset werd een error van 7.40% behaald. Er werd een grafische gebruikersinterface ontwikkeld voor visualisatie en klinisch gebruik van de segmentatieresultaten. Er werd een 3D-segmentatiemethode ontwikkeld voor ultrasoundvolumes van de schildklier, CA en JV. Dit werk effent de weg voor de ontwikkeling van een plannings- en navigatiemethode voor gebruik met RFA voor schildklier nodi.

Contents

3 Introduction

3.1 Thyroid nodule pathology and treatment

Thyroid nodules are common in the adult population, with approximately 50-70% of adults presenting with thyroid nodules on ultrasound imaging. [1, 2]. Of these nodules, 10-20% are symptomatic [3], leading to aesthetic problems as well as difficulties breathing or swallowing [4]. Approximately 90% of all thyroid nodule cases are benign. [5]. Ultrasound is employed for thyroid pathology diagnosis, in which the volume of the thyroid is an important measure [6]. Caliper measurements and the ellipsoid formula are used to determine the volume [7]. More information on diagnosis of thyroid nodules can be found in Section 4.1.

Surgery is the conventional way of treating benign thyroid nodules. Surgery poses the risk of complications. The most common complications are recurrent laryngeal nerve (RLN) palsy, which can be temporary (up to 1.9%) or permanent (up to 0.6%), postoperative hemorrhage, temporary (up to 30%) or permanent (up to 0.4%) hypoparathyroidism, hematomas (up to 0.4%), recurrence of nodules (up to 1.2%), wound infection (up to 0.6%), and hypocalcemia (up to 30%) [8]. Surgery has other disadvantages, including general anesthesia and scar formation [9]. In recent years, minimally invasive thermal ablation techniques, such as radiofrequency ablation (RFA) under ultrasound guidance have become more frequent in the treatment of thyroid nodules [10]. RFA consists of the insertion of an internally cooled needle into the target nodule. The needle is connected to a generator producing an alternating high-frequency current, causing vibration of ions in the tissue, creating heat. The tissue in contact with the exposed tip will undergo thermal injury and coagulative necrosis in the target nodule, leading to shrinkage of the nodule [11].

While RFA is a relatively safe technique, some complications can still occur. Complications during RFA include temporary voice change with a duration longer than 1 month (0.7%) and shorter than one month (0.1%), nodule rupture requiring drainage (0.1%) or requiring conservative treatment (0.3%), Horner syndrome (0.1%), hypothyroidism (0.1%), hematomas (0.8%), hypertension (0.5%) for patients with benign thyroid nodules [12].

However, some studies also reported no major complications nor an affected thyroid function after RFA [9, 13]. Elaborate information about the RFA procedure can be found in Section 4.1.

During the RFA procedure, the radiologist uses 2D ultrasound to visualize the thyroid and guide the RFA needle. During the procedure, gas formation caused by the ablation deteriorates the visibility of the thyroid on the ultrasound image [14]. The 2D visualization and artifacts during and before the procedure restrict the radiologist from fully monitoring the position of the RFA needle and the local vital structures in real-time [15]. Radiologists require considerable experience to target the correct structures without damaging surrounding organs [16, 17].

To improve the RFA procedure, the use of 3D ultrasound is suggested during the procedure. This increases the field of view (FOV) and reduces the need for transducer movement to visualize the entire needle and vital structures [18]. A computer-aided intervention system could provide better insights into needle insertion placement and prevent damage to nearby critical structures. A critical step in creating such a tool is to acquire an accurate (semi-)automatic segmentation of the thyroid and surrounding vital organs.

3.2 Previous research

In literature, the results of semi-automatic and automatic segmentation methods are mentioned. More information about these segmentation methods can be found in Section 4.3. Research has already been performed on the automatic segmentation of the thyroid in ultrasound images using deep-learning algorithms. [19, 20, 21, 22] Older research focused more on 2D ultrasound segmentation and more recent research also segmented the thyroid from 3D ultrasound images. To the author's knowledge, no research has been performed yet on the segmentation of 3D ultrasound data acquired with a matrix transducer. This research will combine the 3D ultrasound sweep dataset used by Krönke et al. with a dataset acquired with a matrix transducer. A comparison of results of this study to results in other research can be found in Section 7.

3.3 Deep learning

Convolutional neural networks are a common type of deep learning architectures in medical image processing, with the advantage of the ability to capture local relations with small convolution filters to learn lower-level

features of the image, such as edges, corners, and textures, and high-level features like shapes and patterns [23]. A U-Net is a fully convolutional network consisting of an encoder and a decoder path. The encoder down-samples and extracts features, to capture contextual information at different scales [24]. Segmentation models based on U-Nets contraction and expansion structure are widely used in medical image segmentation [25].

3.4 Research goal

This study developed a pipeline that can segment the thyroid, jugular vein (JV), and carotid artery (CA) in a 3D ultrasound scan of the neck automatically using a U-Net. The U-Net was trained on a dataset of 3D US images made with the matrix transducer and tracked sweep. Different training strategies are compared, being training in 2D, a majority vote in axial, sagittal, and coronal orientation, and 3D to find which strategy leads to the best results.

The segmentation is of use in treatment planning and navigation of RFA procedures, to provide a clear 3D overview of the structure that needs to be ablated and the structures that need to be avoided. To facilitate its use in the clinic the pipeline is implemented in a Graphical User Interface (GUI). In addition to treatment planning, the segmentation can also be used for thyroid volumetry during diagnosis and follow-up.

3.5 Research questions

The research goal is divided into the following research questions:

1. What is the potential of using the matrix transducer for 3D ultrasound acquisition in segmenting the thyroid, CA and JV using a U-Net?
2. What training strategy will lead to the best results?
3. How can the results be implemented for treatment planning of RFA procedures?
4. How well can the model predict the volume of the thyroid, compared to the ellipsoid formula used in the clinic?

4 Background

4.1 Clinical background

This chapter will introduce the pathology of thyroid nodules and the procedure of RFA.

4.1.1 Thyroid nodule pathology and diagnosis

The thyroid is located anteriorly in the neck across the front of the trachea, composed of a left and a right lobe with a small connecting branch, called the isthmus. The thyroid produces the thyroid hormones thyroxine and triiodothyronine, essential for normal development, growth, and metabolism, and calcitonin, which plays a role in Ca^{2+} and phosphate homeostasis [26].

A frequently occurring pathology in the thyroid is the presence of thyroid nodules, which are caused by an overgrowth of cells in the thyroid gland. The majority of people diagnosed with thyroid nodules are asymptomatic. Those who do experience symptoms can have a globus sensation, difficulty swallowing, shortness of breath, hoarseness and pain [5]. Thyroid nodules can also be functioning autonomously, causing hyperthyroidism over time [27].

Sonography is used as the primary modality for the initial stratification of cancer risk and to decide on the need for a fine-needle aspiration biopsy. Ultrasound is suggested when the thyroid gland is palpably abnormal or upon incidental detection in other radiological studies. Owing to the superficial location of the thyroid, high-resolution ultrasound probes ($\geq$ 12 MHz) can be used to detect the nodules. For the assessment of nodules, the American College of Radiology has recommended a point system called the Thyroid Imaging Reporting and Data System (TIRADS). This system assigns points based on 5 ultrasound features, which determine the estimated cancer risk and recommendations for fine-needle biopsy or surveillance [5].

4.1.2 Radiofrequency ablation

In the past years, minimally invasive techniques under US-guidance have become more frequent in the treatment of thyroid nodules. The guideline of image-guided thyroid ablation in Europe and Asia states that chemical and thermal ablation techniques have been proposed as common modalities for non-surgical treatment of benign thyroid nodules [28]. The current practice guidelines state the use of laser ablation (LA) and RFA as first-choice thermal ablation treatment modalities [28]. Recent studies have shown that RFA showed a larger volume reduction and fewer overall complications than LA, with a smaller number of treatment sessions [28]. Studies have shown that volume reduction in 6 to 12 months can be 50,0 to 93,4% [11]. Despite these good results, a survey of the European Thyroid Association showed that in the management of thyroid nodules only 16% of European Thyroid Association members had availability to thermal ablation procedures and only 5% performed thermal minimally invasive techniques themselves [29]. RFA is a relatively safe technique, but some complications can still occur [3].

The ablation procedure starts with the insertion of an internally cooled electrode into the target nodule. See Figure 1 for a schematic image of the needle placement.

Figure 1: Schematic of the thyroid and surrounding critical structures during an RFA procedure. The CA (red) JV (blue), and nerves (red) are near the ablation zone [30]. The numbers represent the order of ablation.

The patient is placed in a supine position with a hyper extended neck to allow for visualization of the target nodule and vital cervical structures with US in real-time. The patient is kept alert and verbal during the procedure, which is of importance for detection of paralysis of the vocal chord during the procedure and to indicate pain, which can serve as a marker for unwanted damage. To separate the target lesion from the surrounding structures in the neck, such as the CA, RLN, JV, anterior cervical muscles, esophagus, and trachea, a hydrodissection technique can be applied. This provides a safety margin to prevent thermal damage to these critical structures. One approach to insertion of the electrode is the trans-isthmic approach, where the electrode is inserted from the midline of the neck and advanced laterally into the target nodule. This limits heat exposure to the RLN. This does make it more challenging to change the position of the electrode when the patient talks or swallows compared with the lateral to medial approach. This is also called the moving shot technique, where the nodule is ablated bit by bit, starting inferior-posterolaterally and moving medially and anteriorly. The deepest part is then ablated first, followed by the more superior layers. This is then repeated for the middle and superior parts [3, 11].

4.2 Ultrasound

Ultrasound is a non-invasive, real-time imaging modality. This section will give a short introduction to ultrasound, the phenomenon of speckle noise, how the ultrasound waves are generated, and 3D ultrasound.

Medical ultrasound imaging is an imaging modality that generated images using sound waves of frequencies of 1 to 20 MHz, The so-called sonogram is generated with a transducer that sends pulses of ultrasound into the body. The sound propagates through the tissue and generates scattered and reflected waves. The waves that are scattered or reflected to the transducer generate the sonogram. The analog signal is discretized, resulting in a finite number of amplitudes that can be read. The resolution in amplitude is 8 (255 levels) or 16 bits (65535 levels) in common medical systems. If the signal becomes too large, it becomes clipped, so in case of 8 bits, all signals over the 255th level become clipped. Signals that are too low have a low signal to noise ratio.

Ultrasound suffers from speckle noise, which is inherent in ultrasound images. It shows as a granular pattern, resulting from constructive and destructive interference of backscattered ultrasound from scatterers smaller than the spatial resolution of the systems [31]. Speckle noise is a random process, but it does contain information. The statistics of the speckle can provide information about different tissue microstructures, but there is no consensus yet on how this can best be interpreted and used. However, it is known that speckle noise reduces image contrast, and blurs and obscures image details [32].

4.2.1 3D ultrasound

Different types of medical transducers are used: linear arrays, curvilinear arrays, phased arrays and annular arrays. Ultrasound is conventionally used as a 2D imaging modality, but approaches for 3D imaging were also developed. This is done using linear arrays in tracked and mechanical scanning. In tracked scanning, a position sensor, e.g. optical or electromagnetic sensor, is placed on the transducer to measure its position and orientation while it is being moved, creating a stack of 2D images that are reconstructed to a 3D image [33]. In mechanical scanning, a motorized mechanism moves a conventional transducer internally to be able to reconstruct a 3D image from acquired 2D images [34]. Lastly, 3D US transducers consisting of a 2-dimensional array of elements can be used [35]. The different types of transducers used to create 3D US images are visualized in Figure 2.

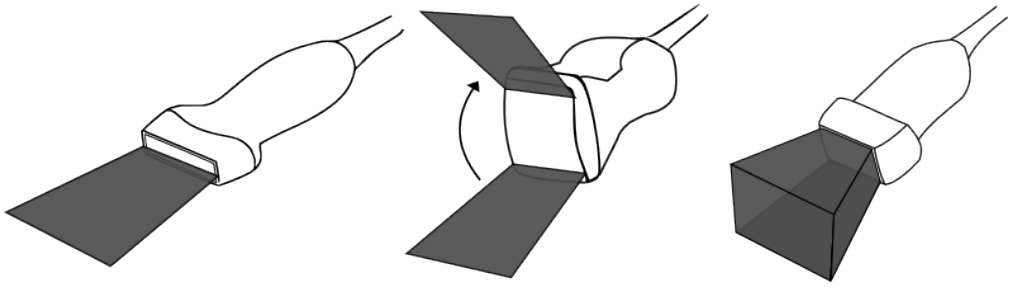


Figure 2: Types of transducers used to make a 3D US scan. 1) A linear 2D transducer that can be swept over the neck, reconstructing 2D scans to a 3D volume. 2) A wobbler transducer where the elements make a mechanical sweep internally while the transducer is being held in place. 3). A matrix transducer, having elements 2-dimensionally, creates a 3D scan in real-time.

The acquired 3D images are usually visualized as multiplanar views or volume rendering. The spatial resolution is usually anisotropic and poorer than in 2D imaging [36].

The tracked method has the advantage over a 3D US transducer of a larger field of view. However, this technique is relatively difficult and localization errors from the sensors can lead to lower accuracy of the resulting calibration [34]. The wobbler transducer has more potential in limiting image distortion, motion and deformation artifacts on both rigid and deformable bodies than the tracked method [34]. The frame rate of the matrix transducer allows for real-time 3D US imaging. Matrix transducers of up to 56.000 elements are available [37].

4.3 Segmentation

The study started with finding a method for segmenting the thyroid. In literature, multiple methods that had already been applied in thyroid segmentation were found. This chapter aims to elaborate on these segmentation methods. Before machine learning was applied in medical image segmentation, the most used approaches were model-based, atlas-based, or a combination. Model-based methods use a ground-truth segmentation mask to build statistical models that capture the shape and appearance of segmentation objects [23]. Atlas-based models register multiple atlases images to the target images, where pixel-wise label predictions are generated with a statistical label fusion, elastic transformation, or another fitting model [23]. The atlas models can be combined with model-based methods as well, for example, shape/appearance models and intensity models [38].

Active contour models, also called snakes, take an initial manually drawn contour and try to optimize this to the actual contours of an object. The initially drawn contour is iteratively deformed to get closer to pixels with a high gradient and to be smooth. For all points on the boundary, it is moved within a certain window where the energy for the contour is minimum. This is iterated until the sum of motions of the contour becomes lower than a certain threshold [39].

Level sets start with an initial guess of the boundary. The boundary curve evolves based on the Gaussian gradient of the image magnitude giving the velocity. The gradient magnitude of the motion of the curve is then calculated. These two are combined in a PDE, that is numerically solved [39].

In graph cut, the image is represented as a graph with the pixels as nodes and the edges represent the similarity of neighboring pixels. The segmentation algorithm tries to find the cut in the graph separating foreground and background regions [39].

In decision tree models the data is split into subsets based on certain features that the algorithm finds, like pixel intensity or texture. Once the algorithm has learned how to partition the data into different categories, pixels from new data can traverse the tree and get assigned a label based on which node it reaches [39].

The advantages of these methods are that it is easy to implement and does not require high-performance hardware devices [4], but it comes with the cost of relatively low performance in comparison to deep learning [38]. Furthermore, deep learning has already been widely studied because of its great performance and results and potential for further improvement in computer-aided diagnosis and computer aided intervention [40].

4.3.1 Deep learning

Modern machine learning segmentation methods currently researched are mostly deep learning models, which outperform the machine learning algorithms and atlas-based auto-segmentation because of the excellent abilities of feature extraction, representation and generalization [23][38]. Deep learning has substantially gained in popularity because of the high-level parallel processing abilities of current hardware and large data availability. The biggest improvement of deep learning as compared to atlas-based methods is seen in segmentation of low-contrast organs [41]. However, class imbalance, which is caused by the large size difference between small and large organs, still can cause worse segmentation results for the smaller organs compared to larger organs [41].

The cost or loss function is a surrogate measure for the performance of the machine learning model, by measuring how well the current output corresponds to the ground truth. The goal of training a neural network is to find the optimal values of parameters for the network by minimizing the cost function. The method of finding these optimal values is determined by the type of optimizer used, with the learning rate serving as the step size for each iteration towards the minimum of the cost function. Regularization is a modification made to the algorithm intended to improve generalization. This can be in the form of putting restrictions on the parameter values, a preference for a simpler model or encoding prior knowledge. These constraints and penalties can lead to an improved performance on the test set by preventing overfitting and forcing constraints on the output. The amount of training samples that are processed by the network in one forward/backward pass is determined by the batch size. A larger batch size gives a more accurate gradient estimation, but it is limited by the memory usage of hardware. The epoch number gives the number of complete passes through a data set [42].

Convolutional neural networks (CNNs) are the most common type of deep learning architecture in medical image processing, with the advantage of the ability to capture local relations with small filters and propagation of dependencies in the short-range by stacking multiple layers [23]. Convolution is a mathematical operation

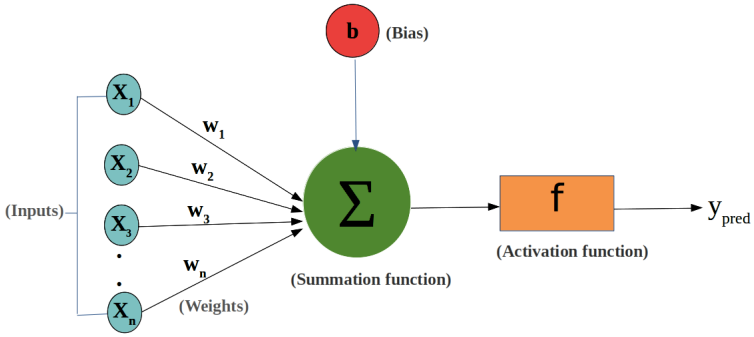


Figure 3: To train a model, it needs to learn what weights to assign to which parameters. These are summed (potentially including a bias), and given to the activation function, which then makes a prediction based on what it has learned.

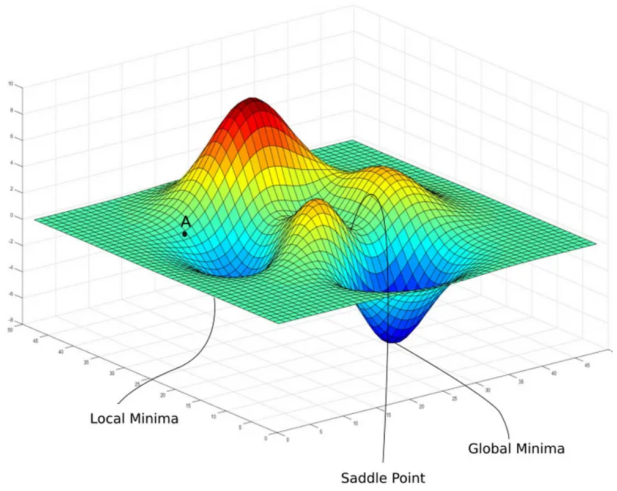


Figure 4: Example of a cost function. To improve results, the model tries to optimize the cost function. The model starts at an initial value of the weight. In each update, a small learning step is taken to find the inverse direction of the gradient. An optimizer is used to find the global minima.

that takes an input (image) and then uses a kernel (filter) to create an output, which is also called a feature map. The multiplication with the kernel results in a smaller amount of features than the original image, for example only edges. CNNs can be used to classify every pixel individually with patches around the pixel and to produce a multi channels likelihood map with the same size as the input image. This will lead to a large memory cost if the dimensions of the feature maps are kept, therefore down-sampling layers such as max pooling and average pooling are applied after some convolutional layers to reduce the dimension of the feature map. However, this does result in a lower resolution than the input image. To prevent this decrease, the fully convolutional network (FCN) can be used [43]. This CNN-based semantic segmentation method replaces the fully connected layers with convolutional layers, to extend the model function from image classification to semantic segmentation. A popular image segmentation network called U-Net stems from FCN. The network has a U-shaped structure with symmetrical encoder and decoder paths. Many variations have already been developed, including UNETR, U-Net++, V-net and deep attention U-Net [23, 44, 45]. More information on the U-Net architecture can be found in Section 5.

5 Methods and materials

A visual summary of the methods is provided in Figure 5. In order to develop a model for segmentation of the thyroid, CA, and the JV, datasets comprised of 3D ultrasound scans of the thyroid needed to be acquired. Two different datasets were used for training and testing. The characteristics of the two datasets are provided below.

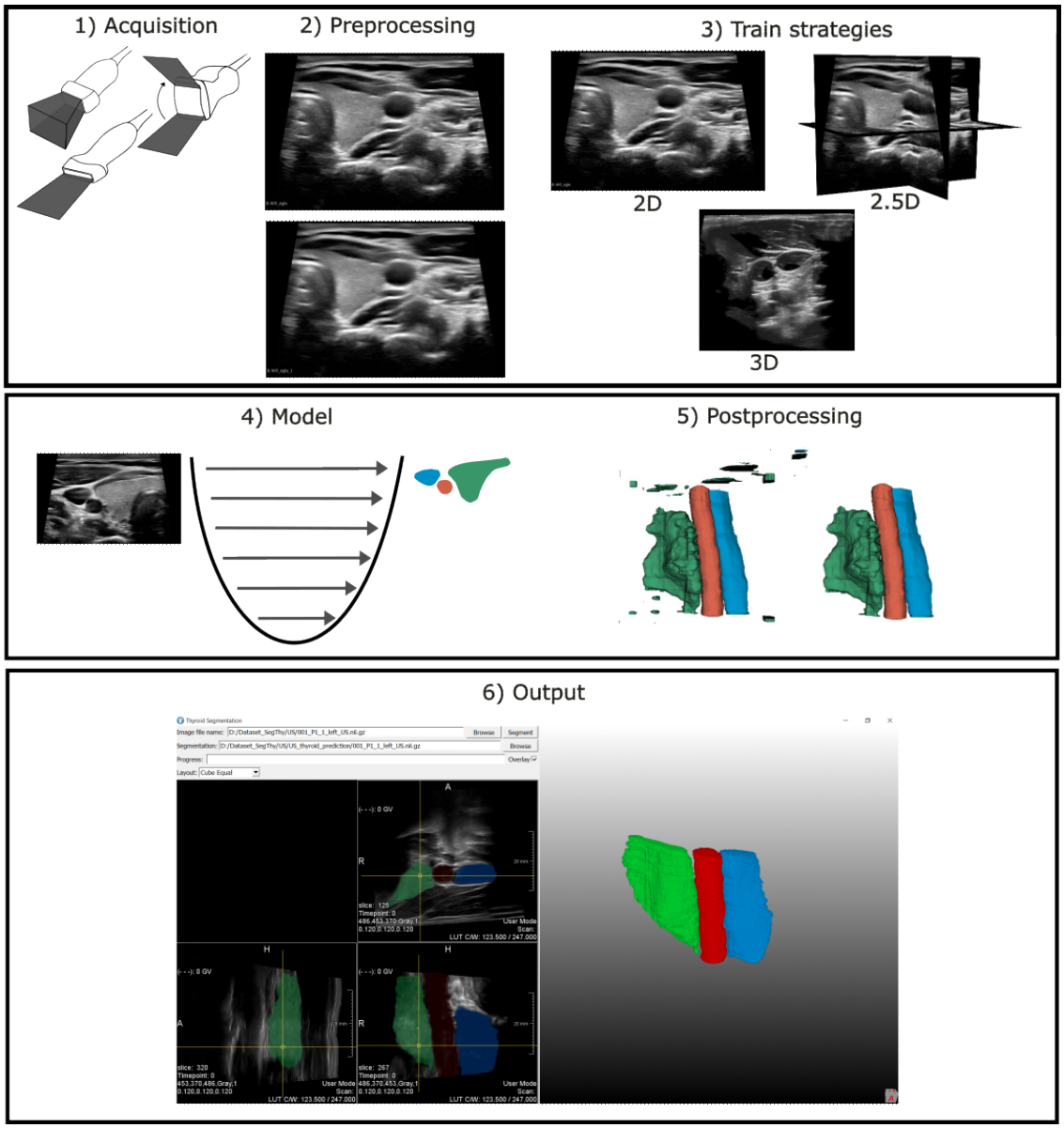


Figure 5: Overview of methods. 1) The model is trained on 3D ultrasound tracked sweep and matrix data. 2) The scans are preprocessed. This involves scaling grey level intensities from 0 to 1 and despeckling in 3D. 3) Different training strategies were used. The first is on 2D axial slices, the second is a majority vote in 2.5D, where slices in all axial, coronal and sagittal planes are used and the third is trained on 3D patches of the scan. 4) The U-Net takes these training strategies as input and produces a segmentation map. 5) The resulting segmentation is post-processed with a largest component analysis. 6) The pipeline is implemented as a GUI in MeVisLab, which visualizes the resulting segmentation as an overlap on orthogonal slices and as a 3D reconstruction for planning of an RFA procedure.

5.1 Data set description

The model was trained on two distinct datasets of 3D ultrasound scans of the thyroid gland. These scans could be acquired using various types of transducers, namely a 2D transducer, a wobbler transducer, or a matrix transducer. To enhance the generalizability of the model across different acquisition methods, the training data set included a publicly available data set obtained from tracked ultrasound scans [21], as well as a self-generated data set acquired using a matrix transducer. The wobbler transducer was also considered, but due to inferior export quality, this was not included in the model.

5.1.1 SegThy tracked sweep dataset

The SegThy dataset [21] comprised tracked ultrasound sweeps of the neck region of 28 healthy volunteers. An example of a sample from this dataset is illustrated in Figure 6. The scans were acquired using a Siemens Acuson NX-3 US machine (Siemens Healthineers AG, Erlangen, Germany), in combination with a 12MHz VF12-4 transducer that employed electromagnetic tracking via a PIUR tUS system (piur imaging GmbH, Vienna, Austria). The scans that were included in the present study were made by a physician with 6 years of experience. The voxel size of the scans was 0.12x0.12x0.12 mm^3 with a variable FOV. Further details can be found in the article by Krönke et al. One scan was excluded due to a lower signal that impeded accurate annotation of the dorsal boundary of the thyroid. The annotations of the thyroid were already created by a radiologist with 8 years of experience, whereas the annotations of CA and JV were added by a Biomedical Engineering master student for this research. The SegThy data set has previously been utilized by Krönke et al. [21] for the development of a deep neural network for thyroid segmentation, with the aim of reducing inter-observer variability in thyroid volumetry. The primary focus was on diagnostic applications, whereas the current study aims to apply the model additionally for treatment planning for RFA procedures.

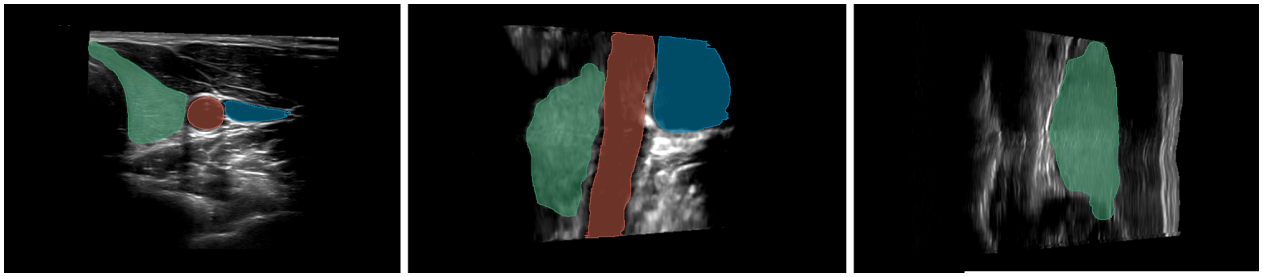


Figure 6: A sample from the SegThy data set, with slices in axial, coronal and sagittal orientation. The thyroid (green), CA (red) and JV (blue) are annotated.

5.1.2 Matrix transducer dataset

The 3D ultrasound data from the matrix transducer dataset was acquired using an XL14-3 xMATRIX transducer connected to an EPIQ Elite 7 ultrasound system (Philips Healthcare, Amsterdam, The Netherlands). A total of 57 volunteers underwent one left and one right thyroid scan. The exclusion criterion was no (partial) removal of a thyroid lobe. The scans were made by a biomedical engineering student. The scans were conducted using the Thyroid protocol of the ultrasound system. The scans have a voxel size of 0.129x0,071x0,141 mm^3 with a fixed FOV. The field of view was set using the maximum angle of 40 degrees. Since this did not capture the entire thyroid, the decision was made only to scan the caudal part of the thyroid. To ensure an equal distribution of samples from both datasets within the final model scans from the first 27 subjects were selected for annotation, with 7 men and 20 women with an average age of 22. An example of a sample from this dataset is illustrated in Figure 7. The dataset was annotated by a Technical Physician and a Biomedical Engineering student and refined by a radiologist with 15 years of experience. Annotations were made in 3D Slicer (version 5.2.1, available at www.slicer.org) [46].

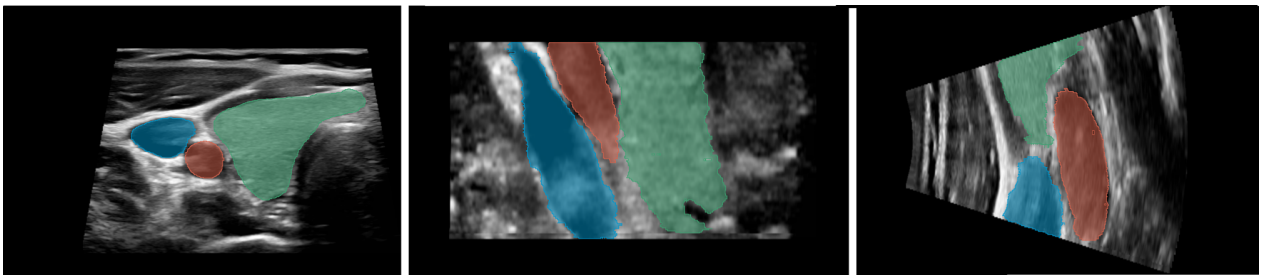


Figure 7: A sample from the matrix data set, with slices in axial, coronal and sagittal orientation. The thyroid (green), CA (red) and JV (blue) are annotated.

5.2 Pre-processing and post-processing

The ultrasound scans underwent pre-processing to enhance the potential of the model segmentations. To obtain equal pixel size for all scans, the voxel spacing was converted to 0.12x0.12x0.12 mm^3. The voxel grey values were scaled from 0 to 1. Ultrasound imaging is susceptible to artifacts and speckle noise, which can hinder segmentation accuracy. Speckle noise is inherent in ultrasound images, caused by microstructures smaller than the imaging resolution. Noise filters are often employed before applying segmentation methods to improve accuracy [47, 31, 48, 6]. A risk of applying a despeckling filter to US images is losing edge information, because of over-smoothing. Many de-speckle methods exist and have been applied to ultrasound images. The best results were found with edge-preserving despeckling methods [47]. The speckle-reducing anisotropic diffusion (SRAD) filter was implemented due to its edge-preserving and enhancing effects. A 3D SRAD filter based on the article by Yu et al. [48] was used to despeckle the images. The Matlab code made by F. Lance [49] was converted to Python code. The manual ROI selection needed for the speckle variation coefficient was replaced with an approximation, as suggested in the article by Yu et al. Additional details regarding this method can be found in Appendix A.1.

Since all final structures consist of one connected part, a keep largest component post-processing step was added to the segmentation pipeline to remove non-connected segmented pixels to improve final results.

5.3 Model description

Because of the excellent segmentation results mentioned in literature, a U-Net architecture was chosen for this application. The U-Net provided by the Medical Open Network for Artificial Intelligence (MONAI) library [50] was used. See Figure 8 for a schematic visualization of the model. This architecture consists of an encoder and a decoder. The encoder consists of convolutional layers with a 3x3 kernel size and a rectified linear unit activation function. This is followed by max-pooling layers with a 2x2 kernel size and stride 2 to downsample the feature maps. The decoder consists of up-convolutional layers with the same kernel size and stride. The corresponding feature maps from the encoder and convolutional layers are then concatenated to generate the final output. This U-Net had 16 to 256 feature channels. A softmax activation function is used in the output layer.

The model was trained with a Dice-cross-entropy loss function. The data was augmented with a random zoom with a factor of 0.8-1.2 during training. The models were trained with a batch size of 8, and an Adam optimizer was used with a learning rate of 1e-3 for the first 400 epochs and 1e-4 for the last 100 epochs, totaling 500 epochs. Lastly, a dropout layer of 0.1 was applied to prevent overfitting of the model.

The model was first optimized on the tracked sweep dataset, because of rapid availability of the data. Afterward, the matrix dataset was added to the training process. Epoch count and learning rate were re-evaluated after combining the datasets. The detailed optimization method is provided in Appendix A.1.

During the optimization phase, 20% of the data was used for validation. The model was tested on 10% of the data and cross-validated 5-fold.

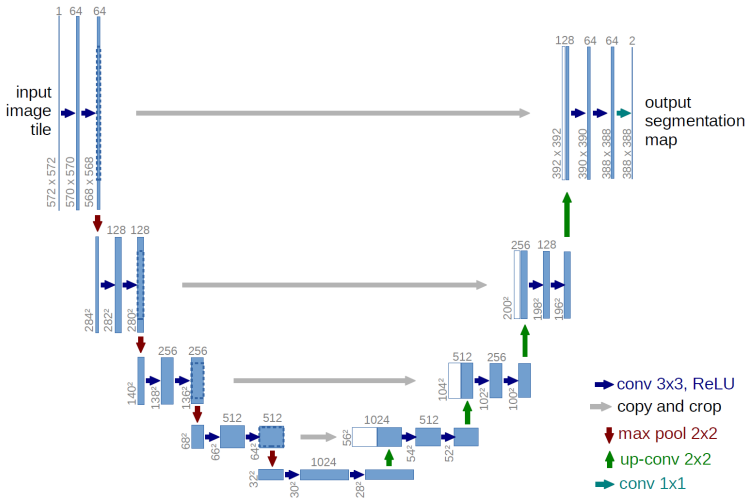


Figure 8: Network structure of a U-Net architecture, consisting of a contracting path and expansive path (encoder and decoder), connected by skip connections. The encoder obtains features from the input image through convolutional operations. Every blue box is a multi-channel feature map and the number of channels is denoted above the box. The number of channels increases while the spatial dimensions decrease by the max-pooling layers. The decoder path uses up-convolutions to recover the original spatial resolution and concatenates the features from the encoder with the upsampled features. [24]. In the U-Net used for this research, the number of channels was changed to 16-256.

5.4 Training strategies

After optimizing the data augmentation and model hyperparameters in 2D, a majority vote assembly in 2.5D, and training in 3D was performed, after which validation loss was reassessed for epoch count and learning rate. For the 2D training strategy, a random axial slice was selected and cropped to 256x256 pixels. The crop was made with 50% of the time the center voxel being a label and 50% of the time, the center voxel being background. For the 2.5D majority vote training strategy, the model was trained on axial, coronal, and sagittal slices separately. For each orientation, a random slice was selected and cropped to 256x256 pixels. The three models were then separately applied to the test samples, and a label was assigned to each pixel only if a minimum of two out of the three models agreed on the label. Finally, for the 3D training strategy, a crop was made of 256x256x64 pixels to make a valid comparison, but limited by memory constraints. The same parameters as used in the 2D model were chosen, but the number of epochs was increased, and the number of non-labeled center crops the model encountered was doubled.

5.5 Evaluation

First, the results of different training strategies were compared. The best performing model in terms of metrics and visual evaluation was used for volume predictions and implementation in the segmentation GUI.

5.5.1 Training strategies

For evaluation, the Dice similarity coefficient (DSC) and Hausdorff Distance 95% (HD_{95}) were used. For testing of differences between the results of different training strategies, a Friedman test ($p = 0.05$) was performed using SPSS (version 28.0, IBM Corp., Armonk, NY, USA) after the cross-validations. A post-hoc Wilcoxon test was conducted if the groups contained significant differences. Additionally, the results of the model were visually evaluated in slice planes and as a 3D reconstruction for their strengths and weaknesses.

5.5.2 Volumetry

For diagnosis and follow-up purposes, the volume of the prediction was compared to that of the ground truth and to the outcomes of using the ellipsoid formula. The ellipsoid formula was only performed on the tracked sweep dataset since the matrix scans do not contain the entire thyroid. Since the isthmus is not included in these measurements in the clinic, a bounding box was made around the thyroid lobes up until the trachea to exclude the isthmus.

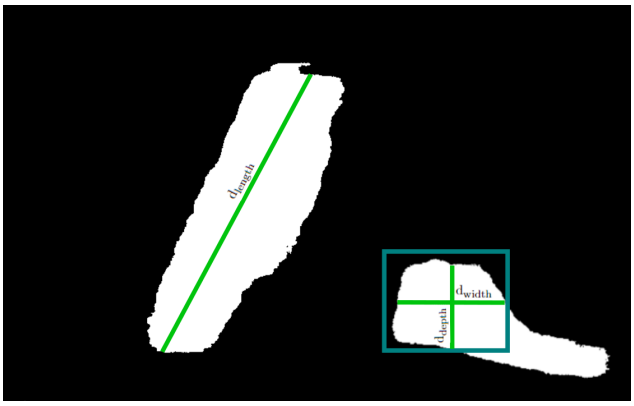


Figure 9: For calculation of the thyroid volume, the ellipsoid formula is used in the clinic. The length, width and depth of the thyroid are manually measured (green lines). For automatic calculations, a bounding box (blue) was made to exclude the isthmus from ground truth and ellipsoid measurements.

To account for inter- and intraobserver variability in drawing the lines representing height, width and thickness of a lobe, the 'regionprops3' function in MATLAB (MathWorks, R2022b) was used to find the length in pixels of the major axes of the thyroid lobe label. This resulted in Equation 1. The results of the matrix transducer volume prediction were compared to the results of the tracked sweep dataset and the results of the prediction of the tracked sweep dataset were compared to the error by using the ellipsoid formula. For the comparison to the model predictions, the best performing training strategies in terms of thyroid DSC and HD_{95} was used. A Mann-Whitney U test was performed to test for statistical differences.

$$V_{ellipsoid} = \frac{\pi}{6} \times V_{pixel} \times d_{length} \times d_{width} \times d_{depth} \tag{1}$$

5.5.3 Segmentation GUI

To demonstrate the practicality of the segmentation pipeline for radiologists, a GUI was developed using MeVisLab (Version 3.4.2, available at www.mevislab.de). The GUI was designed to enable the user to use buttons to load medical images and visualize them in orthogonal planes with the ability to scroll through different slices and in 3D to view results from all angles. The user could visualize and toggle the segmentation results and visually evaluate the accuracy of the segmentation.

6 Results

An overview of the results upon which decisions were made to obtain this model, can be found in Appendix A.2. The first run was performed on a completely held-out test set. Results are shown in Appendix B.

6.1 Training strategies

The performance of the model for segmenting the thyroid, CA, and JV was evaluated on the held-out test set and then cross-validated 5-fold for statistical testing. 5-fold cross validation on 6 samples each fold, led to a total of 30 scans in the test set for both datasets. The results can be found in Tables 1 and 2. The results are also visualized in Figure 10 as a boxplot to visualize the spread and outliers of the metrics of the train strategies. A visualization of results is provided in Figure 11.

There was a difference in the results of the different training strategies for the sweep dataset. As can be seen in Table 1, the DSCs of all structures differ significantly between training strategies, as does the HD_{95} of the JV. A post hoc Wilcoxon test showed that for the thyroid, the DSC was highest in 2D. For the CA, the DSC was highest in 2D and 3D and the HD_{95} was also best in 2D and 3D. For the JV, the DSC was highest in 2D and 3D. P-values of the post hoc Wilcoxon test can be found in Table 3.

There was a difference in the results of the different training strategies for the matrix dataset as well. As can be seen in Table 2, there is a significant difference between the training strategies of the thyroid, CA and JV. post hoc Wilcoxon test showed that for the thyroid, training in 2D and 3D led to the highest DSC. For the CA, training in 2D and 3D led to the highest DSC. For the JV, the DSC was highest for 2D and 3D and the HD_{95} was best when trained in 3D. P-values of the post hoc Wilcoxon test can be found in Table 3.

	DSC			**HD_{95}**		
	T	**CA**	**JV**	**T**	**CA**	**JV**
2D	0.934 ± 0.036	0.924 ± 0.022	0.897 ± 0.112	1.206 ± 1.132	0.588 ± 0.343	1.571 ± 3.255
2.5D	0.920 ± 0.032	0.910 ± 0.043	0.844 ± 0.185	1.368 ± 1.014	1.010 ± 0.556	1.868 ± 2.190
3D	0.917 ± 0.046	0.924 ± 0.030	0.885 ± 0.148	1.397 ± 1.136	0.543 ± 0.450	1.660 ± 3.404
Sig.	<0.001	<0.001	<0.001	0.587	<0.001	0.092

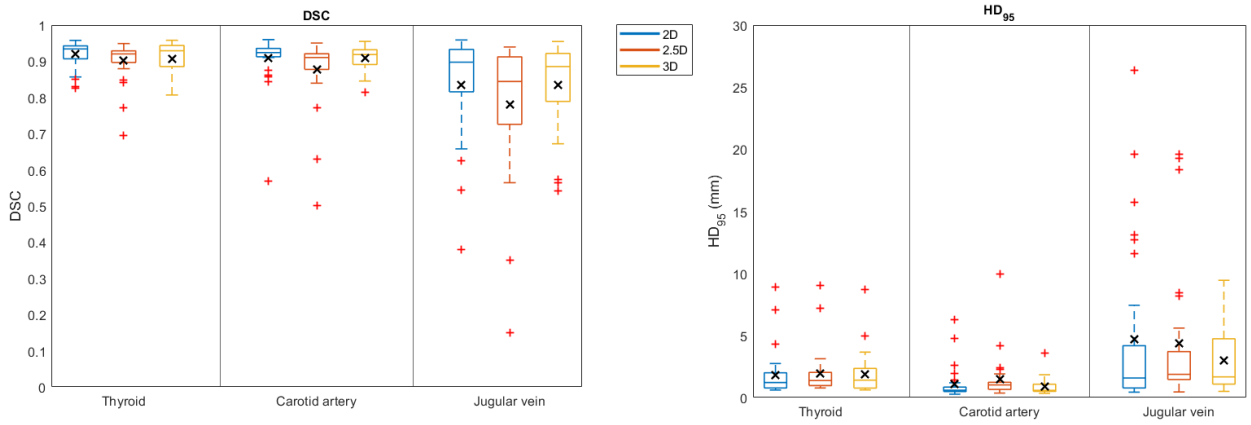
Table 1: Median results tracked sweep dataset measured in DSC and HD_{95} for the thyroid, CA and JV. The results of the Friedman test to test for significant differences between the training strategies are included.

	DSC			**HD_{95}**		
	T	**CA**	**JV**	**T**	**CA**	**JV**
2D	0.894 ± 0.043	0.931 ± 0.041	0.881 ± 0.184	1.712 ± 0.805	0.495 ± 0.911	2.486 ± 3.003
2.5D	0.863 ± 0.074	0.919 ± 0.065	0.824 ± 0.239	1.911 ± 1.080	0.736 ± 1.286	2.301 ± 4.025
3D	0.869 ± 0.045	0.930 ± 0.031	0.856 ± 0.148	1.814 ± 0.618	0.606 ± 0.437	1.405 ± 2.099
Sig.	<0.001	0.009	<0.001	0.113	0.227	<0.001

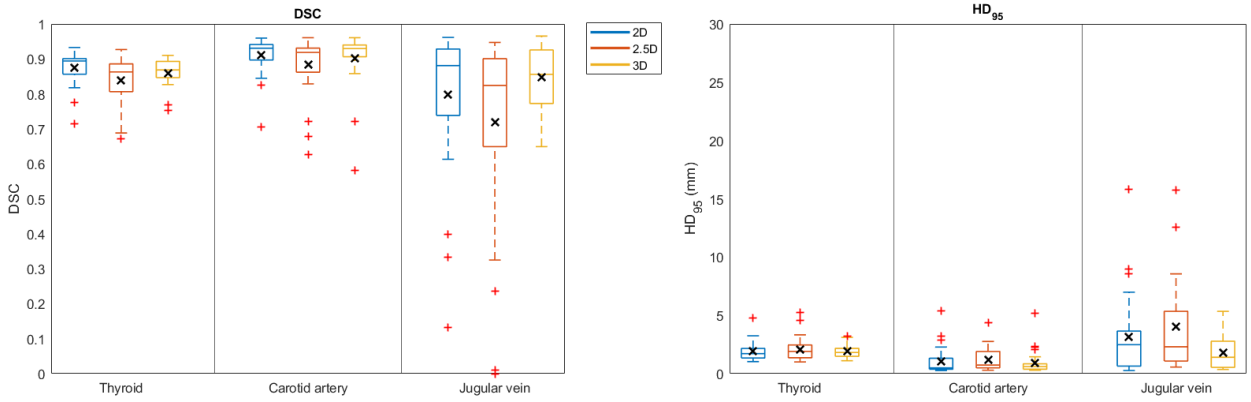
Table 2: Median results matrix dataset measured in DSC and HD_{95} for the thyroid, CA and JV. The results of the Friedman test to test for significant differences between the training strategies are included.

	Sweep				**Matrix**			
	DSC T	**DSC CA**	**DSC JV**	**HD CA**	**DSC T**	**DSC CA**	**DSC JV**	**HD JV**
2D/2.5D	<0.001	<0.001	<0.001	<0.001	<0,001	<0.001	<0.001	0.028
2D/3D	0.043	0.705	0.787	0.234	0.104	0.254	0.144	0.007
2.5D/3D	0.185	<0.001	0.063	<0.001	0.112	0.187	0.004	0.003

Table 3: Significance level of post hoc Wilcoxon test for all groups that showed significant differences in Friedman test.



(a) Tracked sweep dataset



(b) Matrix dataset

Figure 10: Boxplot containing the results of different structures in the tracked sweep and matrix dataset test sets after cross-validations measured in DSC and HD_{95}. Containing the mean (x) and outliers (+).

The results from the overall best performing training strategy of the tracked sweep (2D) is also compared to the overall best performing training strategy of the matrix dataset (3D). Results of the Mann-Whitney U test can be found in Table 4. The DSC and HD_{95} of the thyroid are higher in the tracked sweep dataset.

	DSC			**HD_{95}**		
	T	**CA**	**JV**	**T**	**CA**	**JV**
Sig	<0.001	0.947	0.859	0.004	0.935	0.220

Table 4: Mann-Whitney U test results of comparison of tracked sweep dataset to matrix transducer dataset.

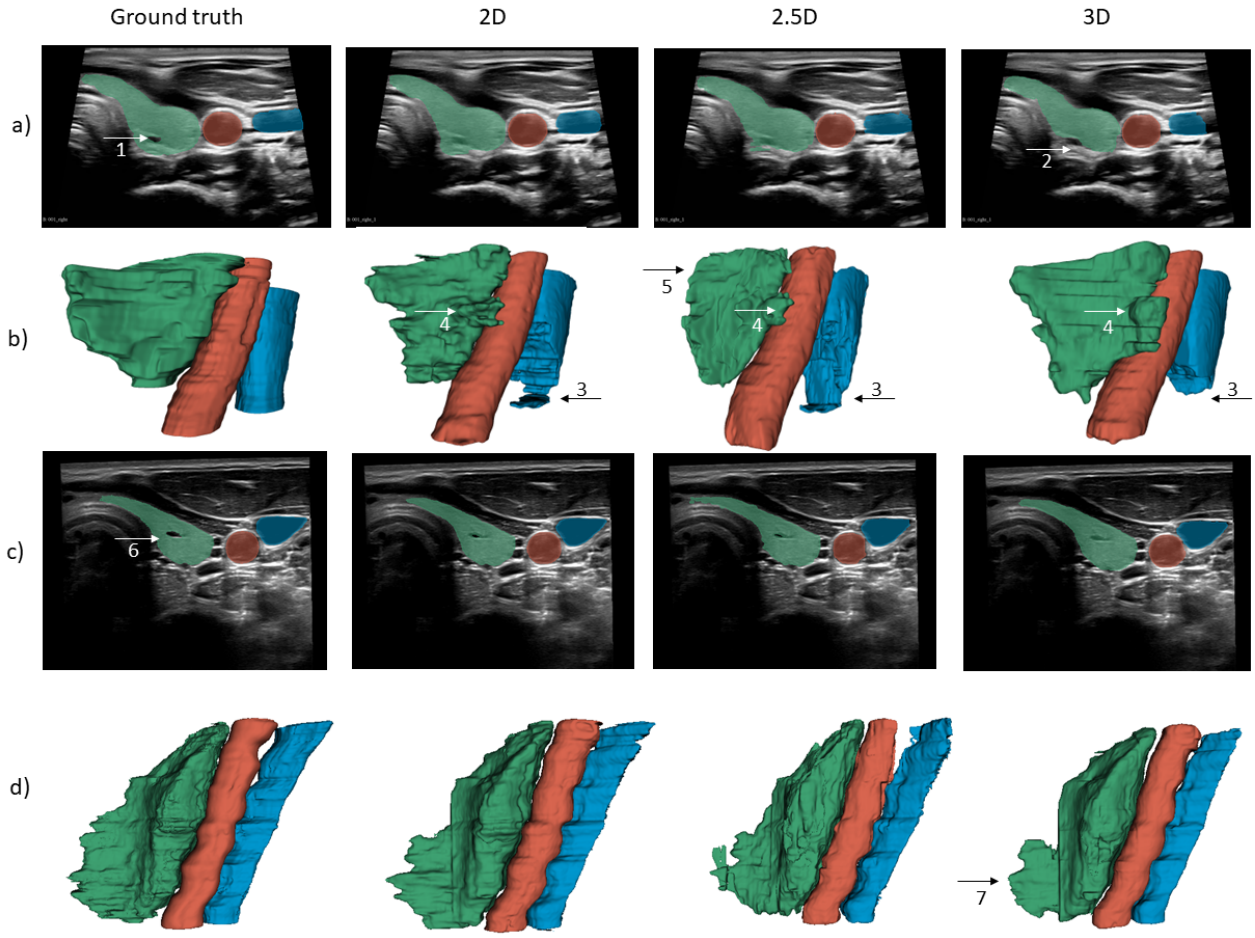


Figure 11: A visualization of two samples from the matrix dataset (a-b) and two samples from the tracked sweep dataset (c-d). a) Axial view: A small vessel traverses through the thyroid (1). In this sample, the 3D model is the only model that excludes the small artery through the thyroid from the thyroid label. However, it misses the distal boundary of the thyroid (2). b) 3D view: All models show some segmentation leakage of the distal part of the thyroid label (4), where the boundaries have a lower contrast with surrounding tissues. The 2.5D model misses the isthmus of the thyroid (5). The JV was not segmented entirely, but was partly labeled as carotid, leading to removal after the postprocessing step. c) Axial view: in this tracked sweep sample, the 2D and 2.5D vessels perform better in segmenting the vessel that traverses through the thyroid than the 3D model (6). d) 3D view: the 3D model creates smoother results than the other models. The isthmus is less well segmented (7). Additionally, this sample visualizes unrealistically large fluctuations in artery and vein dimensions, which are caused by erroneous volume reconstruction of the tracked sweep.

6.2 Volumetry

The volume error was determined with the 2D model, which obtained the best thyroid segmentation results for the tracked sweep dataset and showed equally good results as the 3D model for the matrix dataset. The volume of the prediction segmentation of the tracked sweep dataset had a median error of 4.45% compared to the ground truth annotations. The error of the matrix dataset was 7.40%. Using the ellipsoid formula on the tracked sweep dataset led to an error of 13.84% compared to the ground truth with exclusion of the isthmus. The ellipsoid volume error was higher ($p < 0.001$) than the volume error of the model. The error on the matrix dataset was larger than the error of the tracked sweep dataset ($p = 0.016$).

6.3 Segmentation GUI

To facilitate the use of the model in the clinic, the model was also implemented in MeVisLab with a visualization tool. The interface is shown in Figure 12. The GUI contains a file browser for the 3D US scan. A file, exported from the US system, containing the 3D US scan is needed. Once loaded, the user can press the segment button and the segmentation is executed. Once the segmentation is finished, the segmentation is saved as a NIfTI file to the folder of the scan and can be loaded with the segmentation file browser. The user can visualize the segmentation as an overlay on the orthogonal planes. The user can check the segmentation by scrolling and toggling the overlay. A viewer on the right shows the structures in 3D, which can be viewed from all angles.

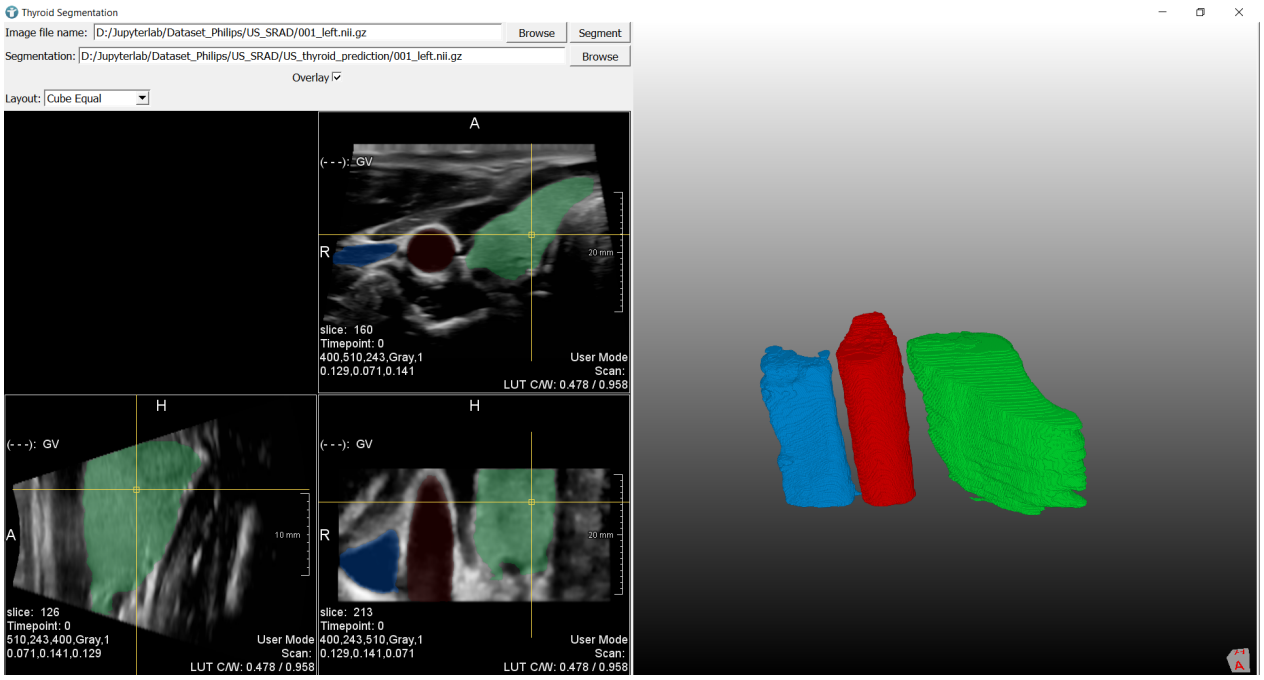


Figure 12: The interface of the segmentation pipeline.

7 Discussion

The goal of this research was to obtain a method that performs 3D segmentation of the thyroid, CA and JV and to present the results in MeVisLab to be used for treatment planning and navigation. Additionally, the segmentation can be used to calculate the volume of the thyroid for diagnostic and follow-up purposes. This study used 3D US thyroid scans made with a tracked sweep and a matrix transducer. A U-Net was applied to segment the thyroid, CA and JV. Different training strategies were applied and the results were cross-validated. The different training strategies had different strengths and weaknesses. The 2D model generally produced the best DSC and HD_{95}. In some cases, small vessels traverse through the thyroid, depending on the orientation of the vessel, the models produce different results. The 2D model generally segmented the isthmus better than the other models. The outliers in the segmentation results were often caused by segmentation leakage to low contrast structures. For the JV, outliers were mostly caused by structures near the edge of the FOV and by mistakes in labeling pixels belonging to the JV as CA. The 2D model was also applied for volumetry of the thyroid. The best prediction was made with the model prediction on the sweep dataset which performed better than the prediction on the matrix dataset and use of the ellipsoid formula.

Of the different training strategies, the 2.5D majority vote performed worst. This was the result of poor segmentation results of training on the coronal planes, which was caused by the low resolution in this plane.

7.1 Comparison to other research

Previous research on thyroid segmentation has already been done. Poudel et al. [51, 19] segmented the thyroid in 2D ultrasound images using semi-automated active contour leading to a DSC of 0.80. Wunderling et al. [52] used semi-automated algorithms, namely level set, graph cut, and decision tree feature classifier, to perform thyroid segmentation. The DSCs had average values of 0.713, 0.748, and 0.601 for the three algorithms, respectively. More information on these methods can be found in Section 4.3. Following up on his previous research on semi-automated algorithms, Poudel et al. [19] created a CNN trained on slices of 3D thyroid scans, leading to a DSC of 0.87. Kumar et al. [20] segmented the thyroid gland, nodules and cystic components on 2D ultrasound using a multi-prong CNN, combining 10 cross-validated models. The algorithm achieved a mean DSC of the thyroid of 0.87 and 0.91 for transverse and longitudinal scans respectively. They compared it to a semi-automated distance regularized level set segmentation, which after placing a seed led to a DSC of 0.94-0.95, thus performed better than their algorithm. Recently, Krönke et al. [21] constructed a convolutional neural network to segment the thyroid in 3D ultrasound images obtained by freehand tracked ultrasound with electromagnetic tracking of a 2D transducer to reduce inter-observer variability of thyroid volumetry. The model was trained on slices. The segmentation resulted in a DSC of 0.95, 0.94, and 0.83 for training, validation, and test set respectively. Ma et al. [22] created a scale-aware attention network and a PointRend technology-Mask R-CNN to segment the thyroid, muscles, trachea, CA, cricoid cartilage, isthmus, esophagus, JV, and endothyroid vessel on 2D ultrasound images. They used mean average precision (mAP) to assess the overlap in segmentations. More information about deep learning algorithms can be found in Section 4.3.1. The results of the best performing model in terms of DSC are in the same range as previous research on thyroid segmentation, but also provide an insight into the potential of 3D ultrasound images acquired with a matrix transducer and the potential use of a deep learning model on 3D ultrasound scans for RFA treatment planning and navigation. The cross-validation led to an amount of 30 samples on which the model was tested, which shows the generalizability of the model to different samples.

Currently, needle positioning and structure visualization are based on 2D US guidance [37]. A 3D segmentation allows for clear structure visualization and may result in more accurate and safe RFA. This research led to a model that can segment the thyroid in both tracked sweep and matrix 3D ultrasound images, while also focusing on the vessels near the thyroid. The segmentation can be used both for diagnosis and for treatment planning.

7.2 Transducers

The different transducer types to make a 3D US scan have different advantages and disadvantages. The 2D transducer with tracking has the advantage of a larger FOV and high resolution. However, longer scanning times of 20-30 seconds or more can introduce motion artifacts. Also, localization errors in the tracker can

reduce scan quality [34]. The wobbler and matrix transducer can create a full 3D volume in less than 3 seconds, reducing the risk of motion artifacts. The limited FOV of the matrix transducer only allows capturing of approximately half the thyroid when making a transverse scan. Because of a higher prevalence of nodules in the caudal part of the thyroid, a decision was made to make the scans with a focus on the caudal part of the thyroid. A longitudinal scan could potentially capture the entire thyroid, but then structures left and right of the thyroid will not be captured. Also multiple transverse scans could be made and stitched together [53]. The wobbler transducer was also considered for this research. However, due to inferior export quality this dataset was not included.

7.3 Limitations

This section provides an overview of limitations of this study.

For this research, healthy subjects were scanned. Diseased thyroids have a different appearance than healthy thyroids, having larger volumes, more irregular shapes and containing large nodules. The trained model might have poor results when applied to patients. To apply the model to patient scans, the model would preferably be trained on a dataset containing patient scans.

The model was trained on two datasets with scans of a combined 72 participants. More datasets or more subjects could be added or acquired to obtain better generalization. Preferably, the scans would be made and annotated by a medical professional, which was not the case for the present study. However, the operator did perform over 100 thyroid-lobe scans already. The thyroid protocol was selected with the automatic optimization setting of the system selected. The current model is trained and tested on both a sweep and matrix dataset. Using different methods of acquisition could already improve generalizability to more different types of acquisition, for example, the mechanically steered wobbler transducer. When applied to scans obtained with other US systems, results could still deteriorate. Other measures taken to increase generalization are preprocessing to equal pixel spacing. Data augmentation was applied to create more variation in thyroid size.

The time the segmentation takes is composed of the preprocessing time and the model application time. The average sample takes 2-5 minutes to be preprocessed. For the 2D model, it takes approximately 0.3 seconds per slice to be applied, leading to 2-3 minutes per 3D volume. The 2.5D model needs to apply three models to the sample, therefore takes about three times longer. The 3D model takes less than 10 seconds to be applied to the entire volume.

The current despeckling method was focused on being automatic, meaning no user input is needed. This led to the decision to choose a standard value for the speckle coefficient of variation q_0. The value was determined by selecting a homogeneous region on the thyroid for all scans in the tracked sweep dataset and calculating the coefficient using Equation 2 for t=0. The value was constant at around 0.7, so this value was chosen. For other US systems however, depending on the amount of speckle development these values might vary.

US has a high inter- and intraobserver variability, because of its low contrast, artifacts and high noise levels [21]. Therefore, creating a ground truth to train the model on is difficult. The model could improve some of the small errors, but these errors will have a negative effect on evaluation results. To potentially find these errors, the models were also visually evaluated on their performance.

The output file of the exported US scan is dependent on the vendor of the US system. The current GUI is made to load NIfTi files. In the clinic, the GUI would have to be tailored to the file that is generated.

7.4 Clinical value

The model could be applied in different clinical settings. The focus of this study was to make a 3D reconstruction that a radiologist can use for planning and navigation of RFA. The radiologist can evaluate the created segmentation by scrolling through the slices with the segmentation as an overlay to the scan. During the procedure, the radiologist can look back on the reconstruction to get an overview of how the thyroid and vessels look in planes that are not visualized or when image quality deteriorates due to gas formation. A patient dataset should be acquired to be able to include nodules as a separate structure in the segmentation. The results could then also be used for needle based diagnosis such as fine-needle aspiration biopsies.

7.5 Suggestions for future research

The current application of the segmentation created in this research is to make a planning pre-treatment. In future applications, research might look into the possibility of an overlay on scans in real-time. The current models are not fast enough to be applied in 3D in real time. Additionally, the quality of the images deteriorates because of the gas bubbles formed. When using a matrix transducer during the ablation, the radiologist could use orthogonal US planes to visualize the thyroid. The scan made pre-treatment could be registered to the US scan in real-time. This does include the assumption that the shape of the structures remains similar during the ablation. The shape of the JV is highly dependent on the amount of pressure applied by the transducer, which would therefore have to be kept as steady as possible.

As mentioned in the introduction, multiple different models have already been applied for thyroid segmentation on 2D and 3D US scans. One of the main features of U-Net is the creation of detailed segmentation maps even with a limited amount of training samples. Also, its context based learning allows for relatively fast training [25]. As mentioned in Section 4.3.1, different versions of the U-Net exist and could potentially create better segmentation results. Examples of models based on the U-Net structure are attention U-Net, inception U-Net, residual U-Net, recurrent U-Net, dense U-Net, U-Net++ and ensemble U-Net [25, 23, 44, 45].

The primary target of the ablation is the thyroid nodule. In the current study, healthy participants were scanned, containing only a few occasional, small nodules, which were now included in the thyroid label. If scans of patients eligible for RFA could be included in the training process, also the nodules could be added as a separate structure for the model to be trained on. As mentioned in Section 4.1, the nerves surrounding the thyroid increase the risk of complications during the procedure. Therefore accurately localizing these nerves, could decrease the risk of complications. However, high variation in the location of these nerves and low visibility make it difficult for an inexperienced person to make these annotations. A safety margin is kept to the esophagus, to prevent esophageal injury, leaving parts of the nodule too close to the esophagus untreated. [54] In future research, the esophagus could also be included in the model.

Furthermore, automatic needle feature localization and tracking could bring improvements to image-guided procedures [55]. Fast and accurate visualization of the spatial relationship between the needle and target can improve the workflow for adjusting the needle position. Furthermore, this information can be used in feedback-controlled robotics-assisted procedures in the future. A needle tracking algorithm could also be used for needle path analysis post-processing to improve procedure planning [55]. Pourtaherian et al. [18] researched the use of a convolutional neural network in needle tracking during 3D US procedures, which led to a higher precision and recall rate as compared to state-of-the-art handcrafted features.

8 Conclusion

For the tracked sweep dataset, overall best results for the test sets were obtained when training in 2D, resulting in a median DSC and HD_{95} of 0.934 and 1.206 mm for the thyroid, 0.924 and 0.588 mm for the CA and 0.897 and 1.571 mm for the JV. For the matrix dataset, the overall best results were obtained when training in 3D, resulting in a median DSC and HD_{95} of 0.869 and 1.814 mm for the thyroid, 0.930 and 0.606 mm for the CA and 0.856 and 1.405 mm for the JV. The tracked sweep dataset gave better results in thyroid segmentation than the matrix dataset in both DSC and HD_{95}, but no differences were found between CA and JV results. The tracked sweep dataset outperformed matrix dataset in thyroid segmentation, but further research is required due to limitations in the matrix transducer's field of view.

The segmentation can give the radiologist an overview of the thyroid, CA and JV in 3D. The overlay in orthogonal planes allow the radiologist to verify the segmentation to determine if the segmentation is accurate enough to use for planning an RFA procedure.

The model can predict the thyroid volume with a smaller error than the ellipsoid formula used in the clinic, with 4.45% for the sweep dataset and 7.40% for the matrix dataset compared to 13.84% with the ellipsoid formula on the sweep dataset.

This research showed the potential of using a tracked sweep and a matrix transducer for 3D ultrasound in segmentation of thyroid ultrasound scans for treatment planning and volumetry, and warrants further research to improve needle-based interventions.

References

[1] Elizabeth H. Holt. Current evaluation of thyroid nodules. *Medical Clinics of North America*, 105:1017–1031, 11 2021.

[2] S. Guth, U. Theune, J. Aberle, A. Galach, and C. M. Bamberger. Very high prevalence of thyroid nodules detected by high frequency (13 mhz) ultrasound examination. *European journal of clinical investigation*, 39:699–706, 8 2009.

[3] Haris Muhammad, Prasanna Santhanam, and Jonathon O. Russell. Radiofrequency ablation and thyroid nodules: updated systematic review. *Endocrine*, 72:619–632, 6 2021.

[4] Dat Tien Nguyen, Jin Kyu Kang, Tuyen Danh Pham, Ganbayar Batchuluun, and Kang Ryoung Park. Ultrasound image-based diagnosis of malignant thyroid nodule using artificial intelligence. *Sensors (Basel, Switzerland)*, 20, 4 2020.

[5] Cosimo Durante, Giorgio Grani, Livia Lamartina, Sebastiano Filetti, Susan J. Mandel, and David S. Cooper. The diagnosis and management of thyroid nodules: A review. *JAMA*, 319:919–924, 3 2018.

[6] Chuan Yu Chang, Yue Fong Lei, Chin Hsiao Tseng, and Shyang Rong Shih. Thyroid segmentation and volume estimation in ultrasound images. *IEEE Transactions on Biomedical Engineering*, 57:1348–1357, 6 2010.

[7] Malcolm C. Brown and Ralph Spencer. Thyroid gland volume estimated by use of ultrasound in addition to scintigraphy. *Acta Oncologica*, 17:337–341, 1978.

[8] Serdar Ozbas, Savas Kocak, Semih Aydintug, Atil Cakmak, Ali Demirkiran, and Gordon C Wishart. Comparison of the complications of subtotal, near total and total thyroidectomy in the surgical management of multinodular goitre. *Endocrine Journal*, 52:199–205, 2005.

[9] Min Ji Hong, Jung Hwan Baek, Young Jun Choi, Jeong Hyun Lee, Hyun Kyung Lim, Young Kee Shong, and Suck Joon Hong. Radiofrequency ablation is a thyroid function–preserving treatment for patients with bilateral benign thyroid nodules. *Journal of Vascular and Interventional Radiology*, 26:55–61, 1 2015.

[10] Mai S. Abd El-Galil, Ali H. Ali, Raef M. Botros, Yasser I. Abd El-Khaleq, and Osama M.A. Hetta. Efficacy and safety of ultrasound (us)-guided radiofrequency ablation of benign thyroid nodules. *Egyptian Journal of Radiology and Nuclear Medicine*, 52:1–11, 12 2021.

[11] Iram Hussain, Fizza Zulfiqar, Xilong Li, Shahzad Ahmad, and Jules Aljammal. Safety and efficacy of radiofrequency ablation of thyroid nodules—expanding treatment options in the united states. *Journal of the Endocrine Society*, 5:1–12, 8 2021.

[12] Cherry Kim, Jeong Hyun Lee, Young Jun Choi, Won Bae Kim, Tae Yon Sung, and Jung Hwan Baek. Complications encountered in ultrasonography-guided radiofrequency ablation of benign thyroid nodules and recurrent thyroid cancers. *European radiology*, 27:3128–3137, 8 2017.

[13] Jung Hwan Baek, Won Jin Moon, Yoon Suk Kim, Jeong Hyun Lee, and Ducky Lee. Radiofrequency ablation for the treatment of autonomously functioning thyroid nodules. *World Journal of Surgery*, 33:1971–1977, 9 2009.

[14] Chiao Yin Wang, Zhuhuang Zhou, Yu Hsuan Chang, Ming Chih Ho, Chiu Min Lu, Chih Horng Wu, and Po Hsiang Tsui. Ultrasound single-phase cbe imaging for monitoring radiofrequency ablation of the liver tumor: A preliminary clinical validation. *Frontiers in Oncology*, 12:3636, 7 2022.

[15] T. Boers, S. J. Braak, M. Versluis, and S. Manohar. Matrix 3d ultrasound-assisted thyroid nodule volume estimation and radiofrequency ablation: a phantom study. *European Radiology Experimental*, 5:1–10, 12 2021.

[16] Wenfeng Song, Shuai Li, Ji Liu, Hong Qin, Bo Zhang, Shuyang Zhang, and Aimin Hao. Multitask cascade convolution neural networks for automatic thyroid nodule detection and recognition. *IEEE journal of biomedical and health informatics*, 23:1215–1224, 5 2019.

[17] Gilles Russ, Adrien Ben Hamou, Sylvain Poirée, Cécile Ghander, Fabrice Ménégaux, Laurence Leenhardt, and Camille Buffet. Learning curve for radiofrequency ablation of benign thyroid nodules. *International journal of hyperthermia : the official journal of European Society for Hyperthermic Oncology, North American Hyperthermia Group*, 38:55–64, 2021.

[18] Arash Pourtaherian, Farhad Ghazvinian Zanjani, Svitlana Zinger, Nenad Mihajlovic, Gary Ng, Hendrikus Korsten, and Peter de With. Improving needle detection in 3d ultrasound using orthogonal-plane convolutional networks. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 10434 LNCS:610–618, 2017.

[19] Prabal Poudel, Alfredo Illanes, Debdoot Sheet, and Michael Friebe. Evaluation of commonly used algorithms for thyroid ultrasound images segmentation and improvement using machine learning approaches. *Journal of Healthcare Engineering*, 2018, 2018.

[20] Viksit Kumar, Jeremy Webb, Adriana Gregory, Duane D. Meixner, John M. Knudsen, Matthew Callstrom, Mostafa Fatemi, and Azra Alizad. Automated segmentation of thyroid nodule, gland, and cystic components from ultrasound images using deep learning. *IEEE access : practical innovations, open solutions*, 8:63482, 2020.

[21] Markus Krönke, Christine Eilers, Desislava Dimova, Melanie Köhler, Gabriel Buschner, Lilit Mirzojan, Lemonia Konstantinidou, Marcus R. Makowski, James Nagarajah, Nassir Navab, Wolfgang Weber, and Thomas Wendler. Tracked 3d ultrasound and deep neural network-based thyroid segmentation reduce interobserver variability in thyroid volumetry. 8 2021.

[22] Laifa Ma, Guanghua Tan, Hongxia Luo, Qing Liao, Shengli Li, and Kenli Li. A novel deep learning framework for automatic recognition of thyroid gland and tissues of neck in ultrasound image. *IEEE Transactions on Circuits and Systems for Video Technology*, 32:6113–6124, 9 2022.

[23] Zhuangzhuang Zhang, Tianyu Zhao, Hiram Gay, Weixiong Zhang, and Baozhou Sun. Weaving attention u-net: A novel hybrid cnn and attention-based method for organs-at-risk segmentation in head and neck ct images. *Medical Physics*, 48:7052–7062, 11 2021.

[24] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. Lncs 9351 - u-net: Convolutional networks for biomedical image segmentation. 2015.

[25] Nahian Siddique, Sidike Paheding, Colin P. Elkin, and Vijay Devabhaktuni. U-net and its variants for medical image segmentation: A review of theory and applications. *IEEE Access*, 2021.

[26] W.F. Boron and E.L. Boulpaep. *Medical Physiology*. Elsevier Health Sciences, 2016.

[27] Roberto Cesareo, Andrea Palermo, Valerio Pasqualini, Silvia Manfrini, Pierpaolo Trimboli, Fulvio Stacul, Bruno Fabris, and Stella Bernardi. Radiofrequency ablation on autonomously functioning thyroid nodules: A critical appraisal and review of the literature. *Frontiers in endocrinology*, 11, 5 2020.

[28] Linye He, Wanjun Zhao, Zijing Xia, Anping Su, Zhihui Li, and Jingqiang Zhu. Comparative efficacy of different ultrasound-guided ablation for the treatment of benign thyroid nodules: Systematic review and network meta-analysis of randomized controlled trials. *PLoS ONE*, 16, 1 2021.

[29] Laszlo Hegedüs, Andrea Frasoldati, Roberto Negro, and Enrico Papini. European thyroid association survey on use of minimally invasive techniques for thyroid nodules. *European thyroid journal*, 9:194–204, 7 2020.

[30] Ji Hoon Shin, Jung Hwan Baek, Eun Ju Ha, and Jeong Hyun Lee. Radiofrequency ablation of thyroid nodules: Basic principles and clinical application. *International Journal of Endocrinology*, 2012, 2012.

[31] Yingtao Zhang, H. D. Cheng, Jiawei Tian, Jianhua Huang, and Xianglong Tang. Fractional subpixel diffusion and fuzzy logic approach for ultrasound speckle reduction. *Pattern Recognition*, 43:2962–2970, 8 2010.

[32] Oleg V. Michailovich and Allen Tannenbaum. Despeckling of medical ultrasound images. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, 53:64–78, 1 2006.

[33] Laurence Mercier, Thomas Langø, Frank Lindseth, and D. Louis Collins. A review of calibration techniques for freehand 3-d ultrasound systems. *Ultrasound in medicine biology*, 31:449–471, 2005.

[34] David Iommi, Johann Hummel, and Michael Lutz Figl. Evaluation of 3d ultrasound for image guidance. *PLoS ONE*, 15, 2020.

[35] Qinghua Huang and Zhaozheng Zeng. A review on real-time 3d ultrasound imaging technology. *BioMed Research International*, 2017, 2017.

[36] Spyretta Golemati and Demosthenes D. Cokkinos. Recent advances in vascular ultrasound imaging technology and their clinical implications. *Ultrasonics*, 119:106599, 2 2022.

[37] Tim Boers, Sicco J. Braak, Nicole E.T. Rikken, Michel Versluis, and Srirang Manohar. Ultrasound imaging in thyroid nodule diagnosis, therapy, and follow-up: Current status and future trends. *Journal of Clinical Ultrasound*, 2023.

[38] Tomaž Vrtovec, Domen Močnik, Primož Strojan, Franjo Pernuš, and Bulat Ibragimov. Auto-segmentation of organs at risk for head and neck radiotherapy planning: From atlas-based to deep learning methods. *Medical Physics*, 47:e929–e950, 9 2020.

[39] Milan Sonka, Vaclav Hlavac, and Roger Boyle. *Image Processing, Analysis, and Machine Vision*. Cengage Learning, 2007.

[40] Wang Ziyang. Deep learning in medical ultrasound image segmentation: a review. 3.

[41] Elias Tappeiner, Samuel Pröll, Karl Fritscher, Martin Welk, and Rainer Schubert. Training of head and neck segmentation networks with shape prior on small datasets. *International Journal of Computer Assisted Radiology and Surgery*, 15:1417–1425, 9 2020.

[42] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[43] Yu Weng, Tianbao Zhou, Yujie Li, and Xiaoyu Qiu. Nas-unet: Neural architecture search for medical image segmentation. *IEEE Access*, 7:44247–44257, 2019.

[44] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R. Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. *Proceedings - 2022 IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022*, pages 1748–1758, 3 2021.

[45] Fausto Milletari, Nassir Navab, and Seyed Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. *Proceedings - 2016 4th International Conference on 3D Vision, 3DV 2016*, pages 565–571, 6 2016.

[46] Andriy Fedorov, Reinhard Beichel, Jayashree Kalpathy-Cramer, Julien Finet, Jean Christophe Fillion-Robin, Sonia Pujol, Christian Bauer, Dominique Jennings, Fiona Fennessy, Milan Sonka, John Buatti, Stephen Aylward, James V. Miller, Steve Pieper, and Ron Kikinis. 3d slicer as an image computing platform for the quantitative imaging network. *Magnetic Resonance Imaging*, 30:1323–1341, 11 2012.

[47] Min Xian, Yingtao Zhang, H. D. Cheng, Fei Xu, Boyu Zhang, and Jianrui Ding. Automatic breast ultrasound image segmentation: A survey. *Pattern Recognition*, 79:340–355, 7 2018.

[48] Yongjian Yu and Scott T. Acton. Speckle reducing anisotropic diffusion. *IEEE Transactions on Image Processing*, 11:1260–1270, 11 2002.

[49] Frank Lance. 3d speckle reducing anisotropic diffusion - file exchange - matlab central, 2023.

[50] Transforms — monai 0.9.0rc1 documentation. https://docs.monai.io/en/latest/transforms.html.

[51] Prabal Poudel, Christian Hansen, Julian Sprung, and Michael Friebe. 3d segmentation of thyroid ultrasound images using active contours. *Current Directions in Biomedical Engineering*, 2:467–470, 9 2016.

[52] T. Wunderling, B. Golla, P. Poudel, C. Arens, M. Friebe, and C. Hansen. Comparison of thyroid segmentation techniques for 3d ultrasound. *https://doi.org/10.1117/12.2254234*, 10133:346–352, 2 2017.

[53] Philipp Seifert, Thomas Winkens, Leonard Knichel, Christian Kühnel, and Martin Freesmeyer. Stitching of 3d ultrasound datasets for the determination of large thyroid volumes - phantom study part ii: mechanically-swept probes. *Medical ultrasonography*, 21:389–398, 2019.

[54] Jung Hwan Baek, Jeong Hyun Lee, Roberto Valcavi, Claudio M. Pacella, Hyunchul Rhim, and Dong Gyu Na. Thermal ablation for benign thyroid nodules: Radiofrequency and laser. *Korean Journal of Radiology*, 12:525–540, 8 2011.

[55] Xinzhou Li, Adam S. Young, Steven S. Raman, David S. Lu, Yu Hsiu Lee, Tsu Chin Tsao, and Holden H. Wu. Automatic needle tracking using mask r-cnn for mri-guided percutaneous interventions. *International Journal of Computer Assisted Radiology and Surgery*, 15:1673–1684, 10 2020.

A Optimization of the model

A.1 Methods

This section contains an overview of the steps taken to optimize the U-Net before applying different training strategies.

A.1.1 General

Initially, the models are trained for 500 epochs. To avoid under or overfitting, the model's validation loss is assessed post-training, which may result in a modification of the epoch count. Additionally, a seed is applied before the application of transformations to ensure reproducibility.

To evaluate increase or decrease of performance, the DSC, HD and HD_{95} are calculated over 12 unseen samples. A Shapiro-Wilk test was performed to test for a normal distribution of the data. If the data follows a normal distribution, a paired t-test is applied. If the data does not follow a normal distribution, a Wilcoxon signed-rank test will be applied to the data. The test samples were also visually analyzed for their improvements.

A.1.2 Loss functions

The model is trained on four different loss functions, specifically dice, dice-cross-entropy, dicefocal and tversky loss. The optimal loss function, as determined by performance on the validation set, is selected for subsequent testing and further model optimization. The diceloss aims to directly optimize the DSC, which is one of the used metrics for assessing the model results, while the dicefocal loss function is noteworthy for its ability to improve the accuracy of challenging misclassified slices, such as caudal slices. By combining Dice loss with Cross Entropy loss, the model also considers pixel classification probabilities, with a focus on maximizing the likelihood of accurate pixel classification. Utilizing the Tversky loss function, with an emphasis on minimizing false negatives, has the potential to improve the model's ability to accurately segment typically missed portions of the image that are more difficult to classify. Post-processing techniques could be applied to eliminate some of the oversegmentations resulting from this approach. Alternatively, utilizing the Tversky loss function to minimize false positives could reduce the occurrence of segmentation leakage.

Loss function	**Parameters**
Dice	*Include_background = False, softmax = True*
Dicefocal	*Include_background = False, softmax = True*
Tversky	*Include_background = False, softmax = True, alpha = 0.3, beta = 0.7*
Tversky	*Include_background = False, softmax = True, alpha = 0.7, beta = 0.3*
DiceCE	*Include_background = False, softmax = True*

Table 5: Parameters used for the different loss functions. Here alpha is the weight of false positives and beta is the weight of the false negatives.

The loss function with the highest mean DSC and HD was used for further testing. A higher priority is given to the thyroid, then the CA and lastly JV. A better HD is prioritized over a better DSC.

A.1.3 Transforming data

Data augmentation is a common technique used to increase the unreliability of models trained on image data. In this study, a dataset of thyroid images is visually analyzed for intensity, size, and shape characteristics. Relevant augmentations are then added to the dataset and their effect on generalizability to unseen data is examined. If results improve, the augmentation is included in the final model. Data can also be preprocessed to improve results. Data can for example be scaled and in the case of US, despeckled.

The learning process of neural networks can become slower when working with larger integer inputs from 0 to 255. To optimise the learning process, pixel values are normalized between 0 and 1.

To introduce more data variability to the model, a positive or negative shift of 10 intensity values with a probability of 0.25 is added to the dataset. Since there is a variation in placement of the transducer on the neck, a slight variation in the angle can be seen in the US images.To create more samples that contain this variation, a random rotation with a probability of 0.25 and a range of 0.10 radials is added. Further, since the size of the thyroid, CA, and JV can vary, a random zoom with a probability of 0.25 and a factor of 0.8-1.2 is added to the dataset. Since no discrimination has to be made by the model between left and right thyroid, a vertical flip with a probability of 0.5 is added. Finally, denoising methods are tested to improve segmentation results. Gaussian smoothing and SRAD are tested. First, a Gaussian smoothing factor of 1.3 is added. Denoising the image could improve segmentation results, leading to fewer holes in the segmentation. However, it could also decrease segmentation quality of edges. To prevent the edge segmentation quality from decreasing, SRAD is implemented. This method should only decrease speckle noise and is supposed to be edge enhancing, because of edge protection methods and reduction of speckle noise. A description of how this method was implemented can be found in the next section.

Transform	**Parameters**
Scale intensity	*minv=0, maxv=1*
Random intensity shift	*prob = 0.25, offsets = 10*
Random rotation	*prob = 0.25, range_z=[0.15,0.15], mode=['bilinear', 'nearest']*
Random zoom	*prob=0.25,min_zoom=0.8, max_zoom=1.2, mode= ['bilinear','nearest']*
Random flip	*prob=0.5, spatial_axis=0*
Gaussian smooth	*image_key = 'img', sigma = 1.3*
Despeckaling	*see SectionA.1.3.1*

Table 6: Parameters used for the different transformations.

A.1.3.1 Despeckaling algorithm This method is based on the article by Yu et al. [48] and the MATLAB code by Frank Lance [49] that also converts it to a 3D process. SRAD was made to reduce speckle noise while preserving image features. Anisotropic diffusion is a widely used technique that smooths an image but preserves edges. The traditional anisotropic diffusion methods often cause blurring or over smoothing. The SRAD method in the paper aims to prevent this by adaptively changing the diffusion coefficients based on local image features.

First, the data is normalized from 0-1 and rounded to three decimals. Next, a speckle coefficient of variation has to be determined. The original method takes as an input a small homogeneous region in the structure, manually determined by the user, and determines the speckle scale function by using Equation 2.

$$q_0(t) = \frac{\sqrt{\mathrm{var}[z(t)]}}{z(t)} \tag{2}$$

To automate this process, an approximation can be used. See Equation 3.

$$q_0(t) \approx q_0 e^{-\rho t} \tag{3}$$

Where ρ is a constant that we take as $\frac{1}{6}$, concluded from experimental and theoretical results. q_0 is the speckle coefficient of variation in the image. This was determined to be approximately 0.7 in the thyroid region for the SegThy sweep dataset. *I will try to verify this by also determining this factor for the matrix dataset.*

Next, the instantaneous coefficient of variation serves as an edge detector. This is calculated using Equation 4.

$$q(x, y, z; t) = \sqrt{\frac{1/2(|\nabla I|I)^2 - 1/6^2(\nabla^2 I/I)^2}{[1 + (1/6)(\nabla^2 I/I)]}} \tag{4}$$

The diffusion coefficient inhibits smoothing near edges and can be calculated with Equation 5.

$$\frac{1}{1 + [q^2(x, y, z; t) - q_0^2(t)]/[q_0^2(t)(1 + g_0^2(t))]} \tag{5}$$

The divergence can be calculated using Equation 6.

$$\begin{aligned} d^n_{i,j,k} = \frac{1}{h^2}[&c^n_{i+1,j,k}(I^n_{i+1,j,k} - I^n_{i,j,k}) + c^n_{i,j,k}(I^n_{i-1,j,k} - I^n_{i,j,k}) \\ &+ c^n_{i,j+1,k}(I^n_{i,j+1,k} - I^n_{i,j,k}) + c^n_{i,j,k}(I^n_{i,j-1,k} - I^n_{i,j,k}) \\ &+ c^n_{i,j,k+1}(I^n_{i,j,k+1} I^n_{i,j,k}) + c^n_{i,j,k}(I^n_{i,j,k-1} - I^n_{i,j,k})] \end{aligned} \tag{6}$$

With these equations combined, each iteration, the image can be updated with Equation 7

$$I^{n+1}_{i,j,k} = I^n_{i,j,k} + \frac{\Delta t}{4} d^n_{i,j,k} \tag{7}$$

The despeckling is done for 50 iterations with a timestep Δt of 0.05.

A.1.4 Hyperparameters

The batch size is a balance between efficient use of computational resources, noisy gradient estimates and overfitting when choosing a lower batch size. On the other hand, overgeneralization and higher computational cost when choosing a higher batch size. The training batch size is varied from 8 to 16 to 32. An Adam optimizer is used. Choosing a smaller learning rate can lead to the model getting stuck in local minima of the loss function. A larger learning rate can prevent the model from reaching the absolute minimum of the loss function. The learning rate is varied from 1e-4 to 1e-3 and also combined. To change the perceptive field, a kernel size of 3 and 5 are compared.

Hyperparameters	**Variations**
Batch size	*8*
	16
	32
Kernel	*3x3*
	5x5
Learning rate	*0.001*
	0.0001
	0.001 for first 400 epochs, then 0.0001 last 100 epochs

Table 7: Values used for variations of hyperparameters.

A.1.5 Regularization

To prevent over fitting, some regularisation methods are applied. A dropout layer randomly sets a certain percentage of neurons in a network layer to zero during training. This way the network learns what unnecessary learned features are and prevents the neurons in the network from relying too extensively on another neuron.

Weight decay regularizes the model by adding a penalty term to the loss function. Because of this penalty, the model is encouraged to use smaller weights to reduce the complexity of the model. Because of this penalty term, the model could generalize better to unseen data.

Batch normalization is a technique that normalizes the input data to each layer of the network. The normalization is done by adjusting and scaling the activations to make the input data more suitable for the subsequent layers of the network.

For regularization, a batch normalization, dropout layer, and weight decay were investigated for their effect.

Regularization methods	**Value**
Dropout	*0.1*
Weight decay	*0.0001*
Batch normalization	

Table 8: Values used for the different regularization methods.

A.1.6 Combining datasets

After optimizing the model, the acquired xMatrix dataset was combined with the SegThy tracked sweep dataset. Assumed is that the doubling of amount of samples will increase results. However, this is done with the assumption that both datasets have a similar distribution and can be learned from interchangeably. To confirm results don't decrease when combining the models, the final model is trained on the SegThy tracked sweep dataset, the xMatrix dataset and a combined dataset.

A.2 Results

This section contains the results during optimizing the models, based on the methods mentioned in Section A.1. The interpretation of the results will follow in the next section. All tables contain mean results of DSC and HD to the thyroid (T), CA, and JV after trying out different loss functions (table 9), transformations (table 10), hyperparameters (table 11), regularizations (table 12) and lastly combining the datasets (13) If the change or addition led to significantly better or worse results, it is noted with a + or - respectively. Table 10 is compared to the results from the DiceCE loss functions without transformations, with a batch size of 16, kernel size of 3x3, and learning rate of 0.001.

Loss function	**Metric**	**T**	**CA**	**JV**	**Mean**
Dice	DSC	0,814	0,881	0,721	0,805
	HD	3,976	1,061	5,205	3,414
DiceFocal	DSC	0,862	0,897	0,825	0,861
	HD	3,397	0,949	1,538	1,961
DiceCE	DSC	0,869	0,897	0,893	0,886
	HD	3,272	0,907	1,464	1,881
Tversky FP	DSC	0,852	0,831	0,790	0,824
	HD	3,628	3,064	2,869	3,187
Tversky FP	DSC	0,836	0,862	0,849	0,849
	HD	3,744	1,119	1,824	2,229

Table 9: Loss functions

Transformation	**Metric**	**T**	**CA**	**JV**	**Mean**
Scale intensity	DSC	0,910+	0,903-	0,875	0,896
	HD	2,803	0,844	1,568+	1,738
Shift intensity	DSC	0,899+	0,883	0,868	0,884
	HD	2,804+	0,920+	1,685	1,803
Random rotation	DSC	0,887	0,883-	0,800-	0,856
	HD	2,686	1,025-	2,209	1,973
Random flip	DSC	0,891	0,889-	0,836-	0,872
	HD	2,912	1,144-	1,885	1,980
Zoom	DSC	0,918+	0,898+	0,871	0,896
	HD	2,402	0,823+	1,497+	1,574
Crop	DSC	0,763-	0,483-	0,036-	0,427
	HD	4,561-	7,485-	6,042-	6,029
Gaussian smooth	DSC	0,796-	0,890-	0,865-	0,850
	HD	3,968-	0,974-	1,571-	2,171
Despeckle	DSC	0,918+	0,902+	0,811	0,877
	HD	2,545	0,841+	2,101	1,829

Table 10: Transformations

Parameter	Value	Metric	T	CA	JV	Mean
Batch size	**8**	**DSC**	0,913+	0,895	0,875	0,894
		HD	2,471+	0,918	1,610	1,667
	16	**DSC**	0,900	0,894	0,899	0,898
		HD	2,734	0,928	1,512	1,725
	32	**DSC**	0,918+	0,898	0,862	0,893
		HD	2,280+	0,855	1,556	1,564
Kernel	**3x3**	**DSC**	0,900	0,894	0,899	0,898
		HD	2,734	0,928	1,512	1,725
	5x5	**DSC**	0,890	0,891	0,797-	0,859
		HD	2,913	0,925	2,064-	1,967
Learning rate	**0.001**	**DSC**	0,900	0,894	0,899	0,898
		HD	2,734	0,928	1,512	1,725
	0.0001	**DSC**	0,929+	0,897	0,871	0,899
		HD	2,242+	0,800	1,530	1,524
	0.001 first 450 epochs, then decrease to 0.0001	**DSC**	0,926+	0,901	0,882	0,903
		HD	2,230+	0,802	1,541	1,524

Table 11: Hyperparameters

Loss function	Metric	T	CA	JV	Mean
Batch normalization	DSC	0,886	0,808-	0,863	0,852
	HD	2,914	1,171-	1,712	1,932
Dropout	DSC	0,910+	0,907+	0,875	0,897
	HD	2,458+	0,794+	1,518	1,590
Weight decay	DSC	0,898	0,867-	0,791	0,852
	HD	2,761	1,116	2,044	1,974

Table 12: Regularizations

Training dataset	Evaluation dataset	Metric	T	CA	JV	Mean
Only trained on single dataset	Sweep	DSC	0,911	0,903	0,873	0,895
		HD	2,457	0,819	1,519	1,598
	Matrix	DSC	0,805	0,872	0,856	0,844
		HD	4,304	0,807	1,723	2,278
Trained on both sweep and matrix dataset	Sweep	DSC	0,912	0,904	0,879+	0,898
		HD	2,380	0,818	1,472	1,557
	Matrix	DSC	0,832+	0,874	0,871	0,859
		HD	3,897+	0,745	1,594	2,079

Table 13: Combining datasets

A.3 Interpretation of results

Based on having the highest DSC and HD for all structures, all further testing is done with the DiceCE loss function.

Scale intensity led to a significant improvement in the DSC of the thyroid and HD of the JV, but a significant decrease in DSC of the CA. Because of the priority in improving thyroid segmentation, scale intensity was included in the final model.

Random intensity shift improved the DSC of the thyroid significantly and the HD of the thyroid , and CA. However, looking at the results, primarily small non-connected false positives were resolved, but there was an increase in significant parts of the thyroid that were not segmented. Evaluating the effect of this transformation on post-processed labels only led to a significant decrease in DSC of the JV. Therefore, this

augmentation was not included in the final model. Random rotation significantly decreased the DSC of CA, and JV and the HD of the CA. Therefore, this augmentation was not included in the final model.

A vertical flip led to a decrease in DSC of the CA, but an improvement in the HD of the thyroid. Visually the results looked worse, since the thyroid edges were often under-segmented. Therefore, this augmentation was not included in the final model.

A random zoom led to a significant improvement in DSC of the thyroid and CA and an improved HD of the CA, and JV. Therefore, this augmentation was included in the final model.

Gaussian smooth led to a significant decrease in performance for all DSC and all HD. Therefore, this augmentation was not included in the final model.

Despeckling led to a significant improvement in the DSC of the thyroid and CA and the HD of the CA. Visual evaluation showed that the despeckled model resulted in a smoother appearance of the segmentation. However, in some slices the despeckled model also entirely missed the JV, in contrast to the same model trained on the original images, where often still a small part of the JV was found, but the result became very pixelated. The despeckling method was included in the final model.

Training in a batch size of 8 and 32 both led to significantly better results than training in batches of 16. Because of a reduced computational cost, the batch size of 8 was included in the final model. The kernel size of 3x3 performed significantly better than the 5x5 kernel. Therefore a 3x3 kernel was chosen for the final model.

Both the learning rate of 0.0001 and the learning rate of 0.001 for the first 450 epochs and 0.0001 for the last 50 epochs led to better results than only using a learning rate of 0.001. In post-processed results, both improved the same scores, but in non-post-processed results, the decreasing learning rate removed more outliers than the the learning rate of 0.0001.

B Held-out test set

	Metric	T	CA	JV
2D	DSC	0.94 ± 0.01	0.94 ± 0.02	0.90 ± 0.03
	HD_{95}	0.89 ± 0.64	0.50 ± 0.27	1.41 ± 0.62
2.5D	DSC	0.92 ± 0.01	0.91 ± 0.02	0.87 ± 0.05
	HD_{95}	1.70 ± 0.74	0.84 ± 0.78	1.52 ± 0.43
3D	DSC	0.92 ± 0.01	0.92 ± 0.02	0.90 ± 0.04
	HD_{95}	1.16 ± 0.50	0.62 ± 0.17	1.60 ± 1.78

Table 14: Results on held-out test set of sweep dataset

	Metric	T	CA	JV
2D	DSC	0.86 ± 0.07	0.85 ± 0.04	0.77 ± 0.15
	HD_{95}	3.02 ± 1.43	3.53 ± 1.16	2.49 ± 1.98
2.5D	DSC	0.83 ± 0.08	0.854 ± 0.04	0.70 ± 0.21
	HD_{95}	3.03 ± 3.12	2.34 ± 1.05	2.39 ± 1.04
3D	DSC	0.82 ± 0.07	0.83 ± 0.08	0.70 ± 0.50
	HD_{95}	3.15 ± 1.29	2.30 ± 1.24	3.57 ± 5.49

Table 15: Results on held-out test set of matrix dataset