

RESEARCH ARTICLE

Tracked 3D ultrasound and deep neural network-based thyroid segmentation reduce interobserver variability in thyroid volumetry

Markus Krönke^{1,2}, Christine Eilers³*, Desislava Dimova³, Melanie Köhler^{3,4}, Gabriel Buschner², Lilit Schweiger², LEMONIA KONSTANTINIDOU³, Marcus Makowski⁵, James Nagarajah⁶, Nassir Navab^{3,7}, Wolfgang Weber², Thomas Wendler³

1 Department of Radiology and Nuclear Medicine, German Heart Center, Technical University of Munich, Munich, Germany, **2** Department of Nuclear Medicine, School of Medicine, Technical University of Munich, Munich, Germany, **3** Chair for Computer Aided Medical Procedures and Augmented Reality, Department of Computer Science, Technical University of Munich, Garching Near Munich, Germany, **4** Medical Faculty, Technical University of Munich, Munich, Germany, **5** Department of Radiology, School of Medicine, Technical University of Munich, Munich, Germany, **6** Nuclear Medicine, Radboud University Medical Center, Nijmegen, The Netherlands, **7** Chair for Computer Aided Medical Procedures, Whiting School of Engineering, Johns Hopkins University, Baltimore, MD, United States of America

* These authors contributed equally to this work.

* christine.eilers@tum.de



OPEN ACCESS

Citation: Krönke M, Eilers C, Dimova D, Köhler M, Buschner G, Schweiger L, et al. (2022) Tracked 3D ultrasound and deep neural network-based thyroid segmentation reduce interobserver variability in thyroid volumetry. PLoS ONE 17(7): e0268550. <https://doi.org/10.1371/journal.pone.0268550>

Editor: Sathishkumar V E, Hanyang University, KOREA, REPUBLIC OF

Received: February 8, 2022

Accepted: May 2, 2022

Published: July 29, 2022

Copyright: © 2022 Krönke et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: The data relevant to this study are available at <https://www.cs.cit.tum.de/camp/publications/segthy-dataset/>.

Funding: This work was partially supported by the EU grant 688279 (EDEN2020). There was no additional external funding received for this study.

Competing interests: The authors have declared that no competing interests exist.

Abstract

Thyroid volumetry is crucial in the diagnosis, treatment, and monitoring of thyroid diseases. However, conventional thyroid volumetry with 2D ultrasound is highly operator-dependent. This study compares 2D and tracked 3D ultrasound with an automatic thyroid segmentation based on a deep neural network regarding inter- and intraobserver variability, time, and accuracy. Volume reference was MRI. 28 healthy volunteers (24–50 a) were scanned with 2D and 3D ultrasound (and by MRI) by three physicians (MD 1, 2, 3) with different experience levels (6, 4, and 1 a). In the 2D scans, the thyroid lobe volumes were calculated with the ellipsoid formula. A convolutional deep neural network (CNN) automatically segmented the 3D thyroid lobes. 26, 6, and 6 random lobe scans were used for training, validation, and testing, respectively. On MRI (T1 VIBE sequence) the thyroid was manually segmented by an experienced MD. MRI thyroid volumes ranged from 2.8 to 16.7 ml (mean 7.4, SD 3.05). The CNN was trained to obtain an average Dice score of 0.94. The interobserver variability comparing two MDs showed mean differences for 2D and 3D respectively of 0.58 to 0.52 ml (MD1 vs. 2), –1.33 to –0.17 ml (MD1 vs. 3) and –1.89 to –0.70 ml (MD2 vs. 3). Paired samples t-tests showed significant differences for 2D ($p = .140$, $p = .002$ and $p = .002$) and none for 3D ($p = .176$, $p = .722$ and $p = .057$). Intraobserver variability was similar for 2D and 3D ultrasound. Comparison of ultrasound volumes and MRI volumes showed a significant difference for the 2D volumetry of all MDs ($p = .002$, $p = .009$, $p < .001$), and no significant difference for 3D ultrasound ($p = .292$, $p = .686$, $p = 0.091$). Acquisition time was significantly shorter for 3D ultrasound. Tracked 3D ultrasound combined with a CNN segmentation significantly reduces interobserver variability in thyroid volumetry and increases the accuracy of the measurements with shorter acquisition times.

Introduction

Exact thyroid volumetry plays an essential role in monitoring and treatment of many thyroid diseases, especially for the radioiodine therapy (RIT) in hyperthyroidism: The needed activity, commonly calculated using Marinelli's formula, is proportional to the thyroid mass [1]. The latter is obtained using the thyroid volume measurements from two-dimensional ultrasound (2D US) using the ellipsoid formula and a correction factor [2]. The probability of a second treatment after treatment failure is dependent on the dose concept, thyroid volume, and achieved organ dose [3]. Therefore, underestimating the thyroid volume could lead to an underestimated administered dose and thus to the need for a second RIT. On the other hand, overdosage can lead to hypothyroidism and radiation-induced malignancies [4]. Volumetry is also important in diagnosing different thyroid diseases. For example, the thyroid volume is considered when interpreting altered values of TSH (thyroid-stimulating hormone). Currently, 2D US is the standard procedure for thyroid volumetry despite its high inter- and intraobserver variability. Three-dimensional (3D) US was first introduced in the 1970s, and it is believed to be superior to 2D US in terms of user dependency [5]. Similarly, the automatic compounding of 2D US sequences to generate 3D US has been shown to reduce variability [6]. However, the potential advantages of 3D US for volumetry only can be exploited clinically in combination with automatic segmentation. Only in this way 3D US can keep up the speed and ease of use of 2D US employing the ellipsoid formula. Andermann et al. [7] compared volume measurements in 2D and 3D US on healthy volunteers. In the 3D case, the volume was estimated by a multiplanar approximation. However, this 3D approach does not use machine learning and cannot be easily automatized. In another work, Chang et al. [8] developed a thyroid segmentation algorithm for US images based on a radial basis function neural network and a region growing method for shape recovery. Afterwards, a population-based stochastic optimization technique was used to estimate the thyroid volume. No 3D compounding was used. Poudel et al. [9] compared different non-automatic and machine-learning-based segmentation algorithms for freehand 3D US of healthy volunteers. In this work, automatic approaches show better results compared to non-automatic ones. Furthermore, the same group around Friebe et al. analyzed several machine learning and image enhancing methods for thyroid segmentation in US [10–13]. These works do not evaluate, yet, the impact of intra- and inter-observer variability on volumetry. Another automatic segmentation has been developed by Webb et al. [14]. They created a network for the segmentation of nodules, thyroid, and cysts on US thyroid cineclips. Their work, however, focused on the network development and its accuracy with respect to manually segmented 2D ground truth.

In contrast, we present an analysis on volumetry and variability in a realistic setup using 3D compounded ultrasound volumes acquired with a tracked ultrasound system. Analyzing intra- and interobserver variability in thyroid volumetry and comparing 2D and 3D US in this task is a fairly well-analyzed topic in literature [5, 6, 15–23]. However, the combination of 3D US and machine learning for volumetry currently lacks exploration. In this paper, we present a user dependency study on thyroid volumetry with 2D and 3D US in employing a fully automatic segmentation of thyroid lobes by a deep neural network. We aim to analyze whether the hypothesis that 3D US imaging combined with this deep neural network is superior to conventional 2D US in terms of accuracy and variability of thyroid volume estimations is valid. Intra- and interobserver variability was calculated for both modalities and then compared with each other. The accuracy was estimated by comparing both US methods to Magnetic Resonance Imaging (MRI). In this paper, our contributions are:

- We introduce a fully automatic method for 3D volumetry of thyroid lobes using tracked 3D US.

- We comprehensively evaluate our method compared to 2D US by analyzing intra- and inter-observer variability and accuracy by comparing to MRI.
- We evaluate the acquisition time performance of our method compared to 2D US.
- We provide a novel tracked 3D US thyroid dataset as the first bigger and labelled open source dataset that will become publicly available upon acceptance.

Materials and methods

Materials

This study was conducted in the scope of an Institutional Review Board Approval from the Ethical Commission of the Technical University of Munich (approved on April 1st, 2020; reference number 244/19 S). All volunteers gave written informed consent to the US and MRI scans. Eligible participants had to be anamnesticly healthy and at least 18 years old and were recruited from the hospital and research lab environment. Exclusion criteria for possible candidates were subtotal or total thyroidectomy, distinct postoperative or inherent changes in the neck anatomy and MRI contraindications (i.e. claustrophobia, neurostimulators, pacemaker, cochlear implants, medication pumps and shrapnels). Twenty-eight healthy volunteers, aged 23 to 50, eighteen males and ten females, participated in the study. One high-resolution T1 VIBE (Volumetric interpolated breath-hold examination) MRI scan of the neck was acquired with a Biograph mMR machine (Siemens Healthineers AG, Erlangen, Germany) with a field strength of 3 T. The resolution of the MRI image was $0.625 \times 0.625 \times 1.00$ mm³, and the field of view $320 \times 320 \times 80$ mm (Fig 1). The accuracy of the MRI segmentation was evaluated with a thyroid phantom (thyroid ultrasound training phantom, model 074, CIRS). According to the manufacturer, the thyroid volume of each phantom has slight variations of ± 3 to 5 cc but is close to 30 cc. From the MRI scan segmentation, we retrieved a volume of 26.33 ml. US images were obtained with an ACUSON NX3 machine (Siemens Healthineers AG, Erlangen, Germany) using a linear 12 MHz VF12-4 probe (Fig 2). The US images were acquired with a

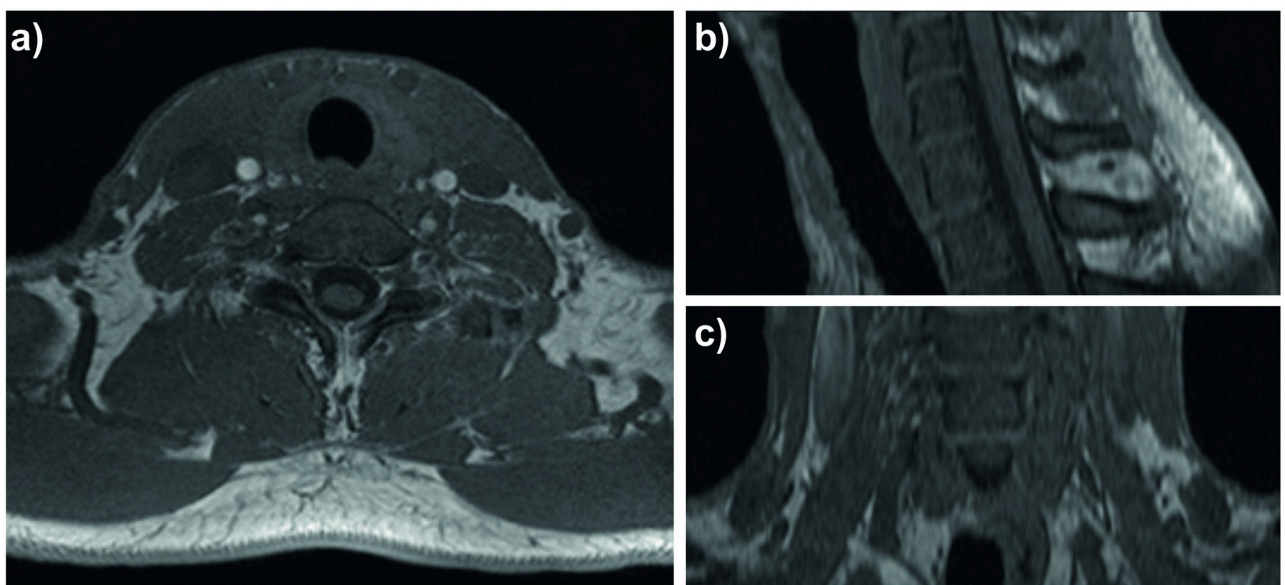


Fig 1. T1 VIBE MRI of the neck region for an exemplary volunteer, a) axial view, b) sagittal view, c) coronal view.

<https://doi.org/10.1371/journal.pone.0268550.g001>

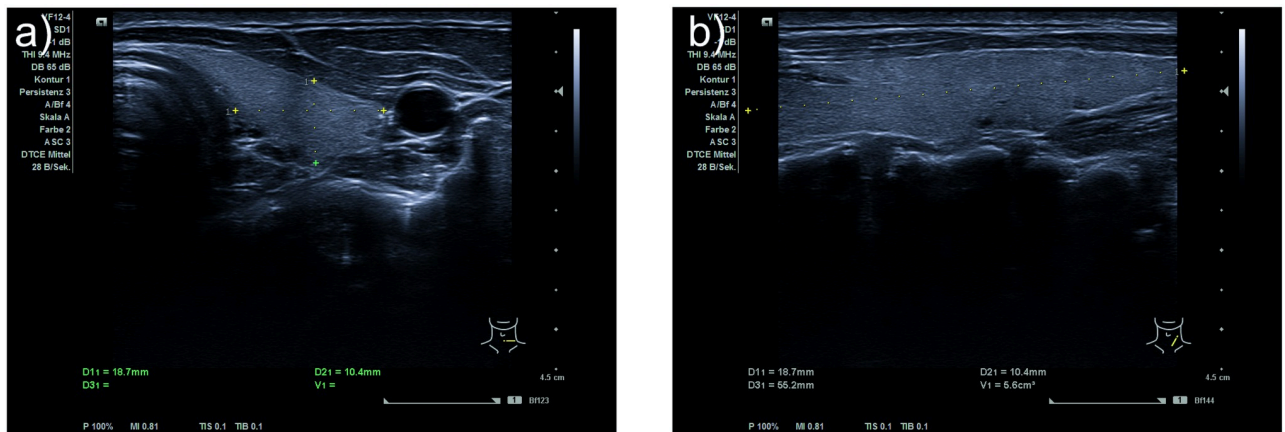


Fig 2. 2D US scans of the neck region for an exemplary volunteer. a) Width and depth of right thyroid lobe marked, b) Length of right thyroid lobe marked.

<https://doi.org/10.1371/journal.pone.0268550.g002>

resolution of 668×599 and a spacing of $0.0753 \times 0.0751 \text{ mm}$. For the 3D US, electromagnetic tracking, and video acquisition, a PIUR tUS system was used (piur imaging GmbH, Vienna, Austria). This device employs two 6D tracking sensors, operating at a frequency of 80 Hz with accuracy in position of 1.40 mm RMS and in angle of 0.50° RMS. Video sequences were acquired with 89 fps. 3D US probes, which consist of matrix arrays, can also be used to obtain US volumes. However, these probes are not widely available, rather bulky, and often do not cover the entire thyroid making a compounding necessary. The image processing (3D US compounding using image data and tracking information) and thyroid segmentation was done using ImFusion Suite Version 2.9.7 (ImFusion GmbH, Munich, Germany). PyTorch was used for the implementation of a deep neural network for automatic 3D US thyroid segmentation and Tensorflow was used for the evaluation of the network. The model was trained on an NVIDIA Titan V 12 GB HBM2 using Polyaxon to schedule different runs (<https://polyaxon.com/>). The training time was around 10 hours in average. The inference on new volumes however took only few seconds. The segmentation pipeline was implemented and trained by our team and differed from the one offered by piur imaging in their latest release. We did not use piur's segmentation tool for this study to be able to modify the segmentation pipeline for our project and to be able to provide this as work open-source for the research community.

General approach

Three 2D and three 3D US scans for each volunteer were acquired by three physicians (MD1, MD2, and MD3) with different levels of experience (6, 4, and 1 years). The ellipsoid formula (correction factor 0.48) was applied to estimate the thyroid volumes from the 2D US scans. No isthmus correction was included. To acquire a volumetric image for a 3D US, the physician slowly sweeps the transducer over both lobes (one sweep per lobe). These sweeps are automatically compounded into a 3D US stack. The 3D US thyroid volumes were calculated based on the automatic segmentation. The resolution of the segmentation mask matched the one of the US image. As a result, volume calculation succeeded by counting all voxels labeled as thyroid. The workflow for the 3D volumetry task can be seen in Fig 3. The thyroid volumes, derived after manual segmentation of the MRI scans by an experienced MD (8 years, cross-sectional imaging), were used as reference.

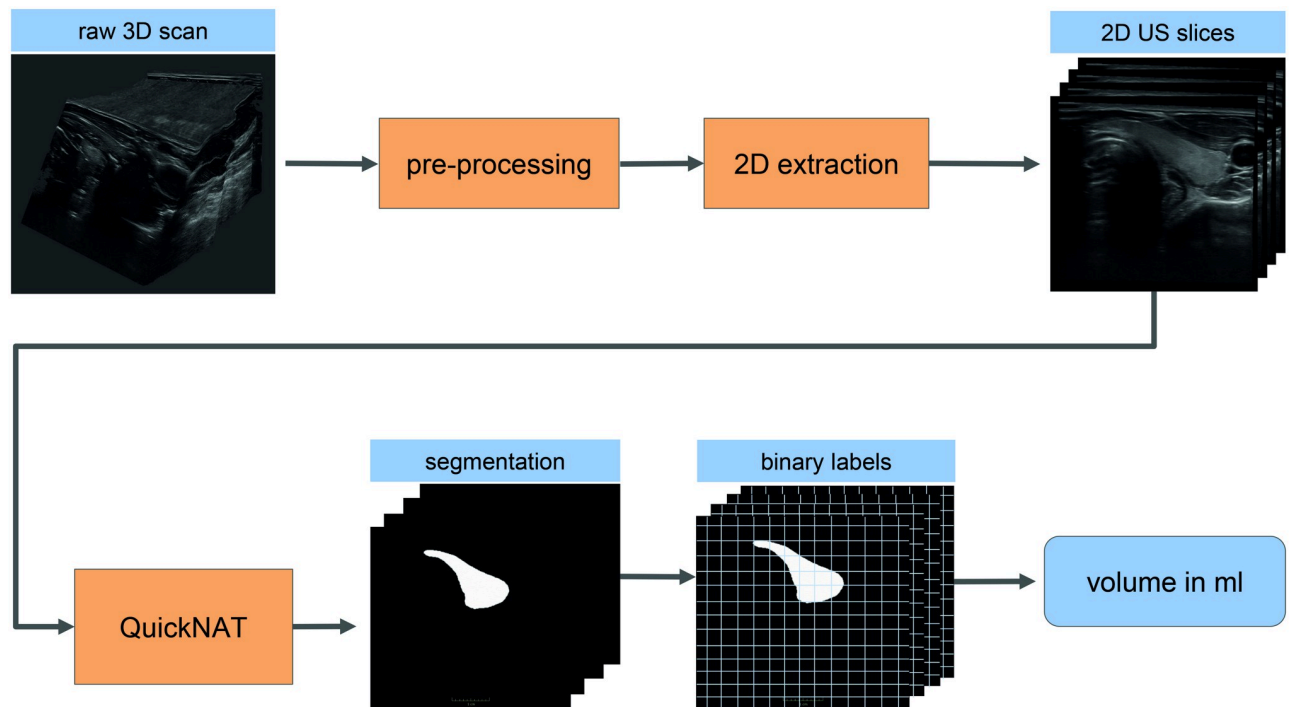


Fig 3. Schematic representation of the automatic segmentation workflow. The raw 3D scans of each lobe are pre-processed by rotation into axial view and resizing as well as centering. Then 2D images are extracted which serve as input to the network. The output of the network is a segmentation from which the volume is calculated by multiplication of pixels which include a segmentation with the pixel volume.

<https://doi.org/10.1371/journal.pone.0268550.g003>

Deep neural network for automatic segmentation

For the automatic segmentation of the 3D US scans, a fully convolutional deep neural network developed by Roy et al. [24] named QuickNAT was used (Fig 4). The structure is similar to the U-net [25], although it can work with smaller datasets as well. Segmentation was done in 2D slice per slice of the compounded 3D volume. QuickNAT consists of dense blocks of four encoders and four decoders with a bottleneck layer in between, as depicted in Fig 4. Each dense block consists of three convolutional layers, two of 5x5 and one of 1x1 kernel size, preceded by a batch normalization layer and a Rectifier Linear Unit (ReLU). The bottleneck block of a 5x5 convolutional layer, including a batch normalization layer, restricts information flow in between the encoder and decoder. There are skip connections between the corresponding encoder and decoder blocks to provide encoder feature information to the decoders directly and improve training through a path of gradient flow to the deeper layers. Each encoder is followed by a 2x2 max-pooling block, which reduces the spatial dimension of the feature maps. To achieve better segmentation for even smaller structures, there are un-pooling layers between the decoder blocks as the indices corresponding to the maximum activations are passed to the decoders. The final layer is a classifier block with softmax, which is a convolutional layer of 1x1 kernel size. It maps the input to a feature map with a given number of channels, two in this case, thyroid and background. Subsequently, the softmax layer transforms the channels into probability maps for each class. Due to the size of our dataset, we evaluated a smaller version of the network with one less encoder and decoder. However, this structural change did not show any improvement in the results. The loss function of the network is a combination of Dice and cross-entropy losses. 26, 6, and 6 random lobe scans were used to train the network, the validation, and testing, respectively. The thyroid in these 3D US scans

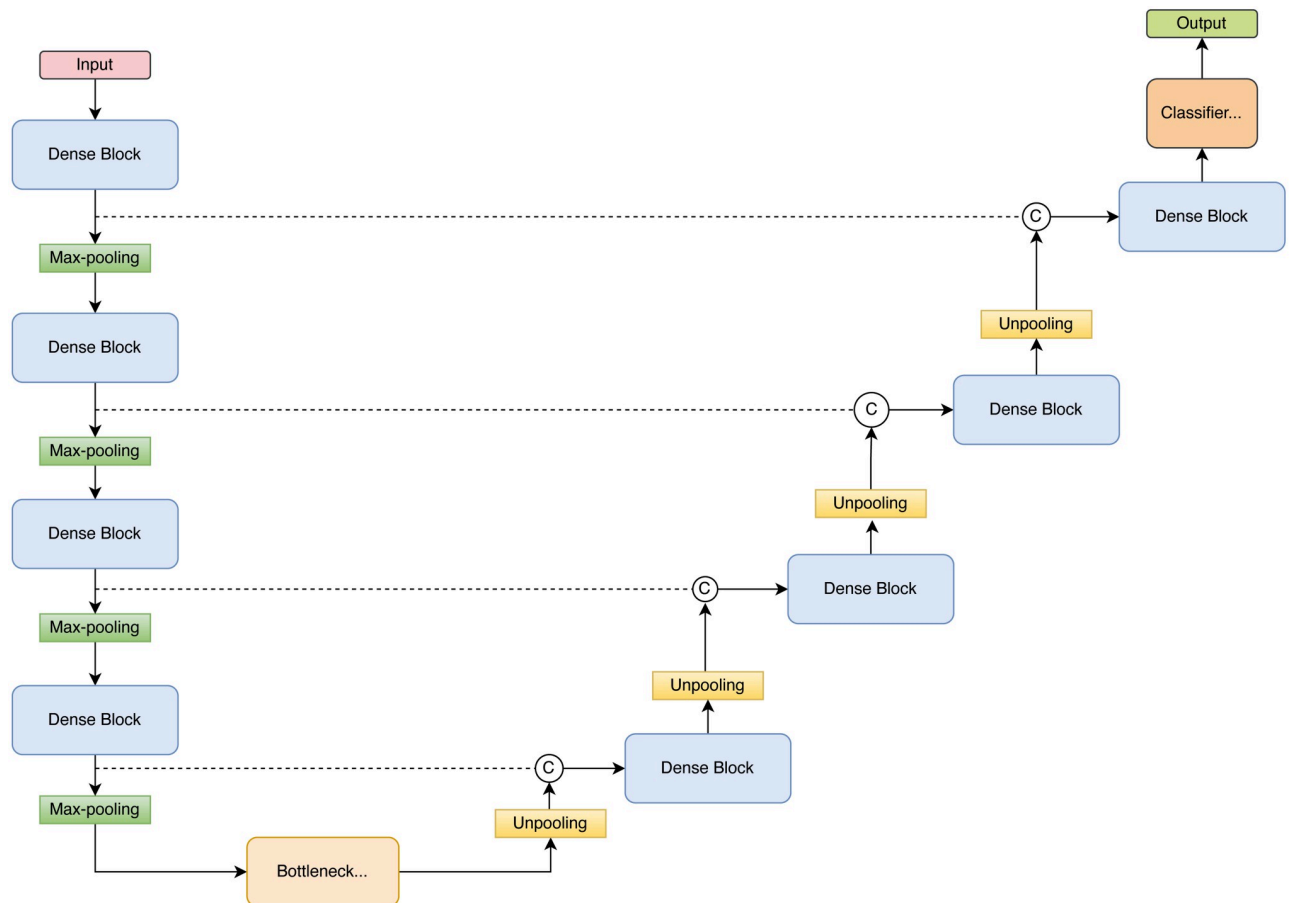


Fig 4. Illustration of the QuickNAT architecture used here for automatically segmenting thyroid in 3D US volumes.

<https://doi.org/10.1371/journal.pone.0268550.g004>

was manually segmented, slice by slice, by an experienced MD (8 years, cross-sectional imaging). Both lobes of the same volunteer were segmented manually which provided a total of 17.664 slices for training (i.e. in average 552 slices per lobe). This total was split into a training set of 14.352 (13 patients) and a validation set of 3.312 images (3 patients). The original architecture of QuickNAT was kept unchanged. A dropout of 0.5, as well as data augmentation, namely random vertical flips and elastic deformations [26], were applied to prevent overfitting. The weights on the Dice and cross-entropy losses were kept equal. Edge weights were used to improve the contour of the segmentation. The number of training epochs was chosen during preliminary experiments, where the Dice loss converged at 20 epochs or earlier. The network was trained with a batch size of 4 at a learning rate of $1e-5$. Dice score amounted to 0.95, 0.94, and 0.83 for training, validation, and test sets. The model was applied to the remaining data (13 volunteers), and the segmentations were used for the volume evaluations (Fig 5).

Statistics and evaluation

From the complete data set, scans from 16 volunteers were used for training and validation. Thus, the network was trained on images that were segmented by the same network in a later step for the volumetry estimation. We, therefore, split the evaluation into two groups, one including scans that were fed to the network during training (15 volunteers) (Group 1) and the

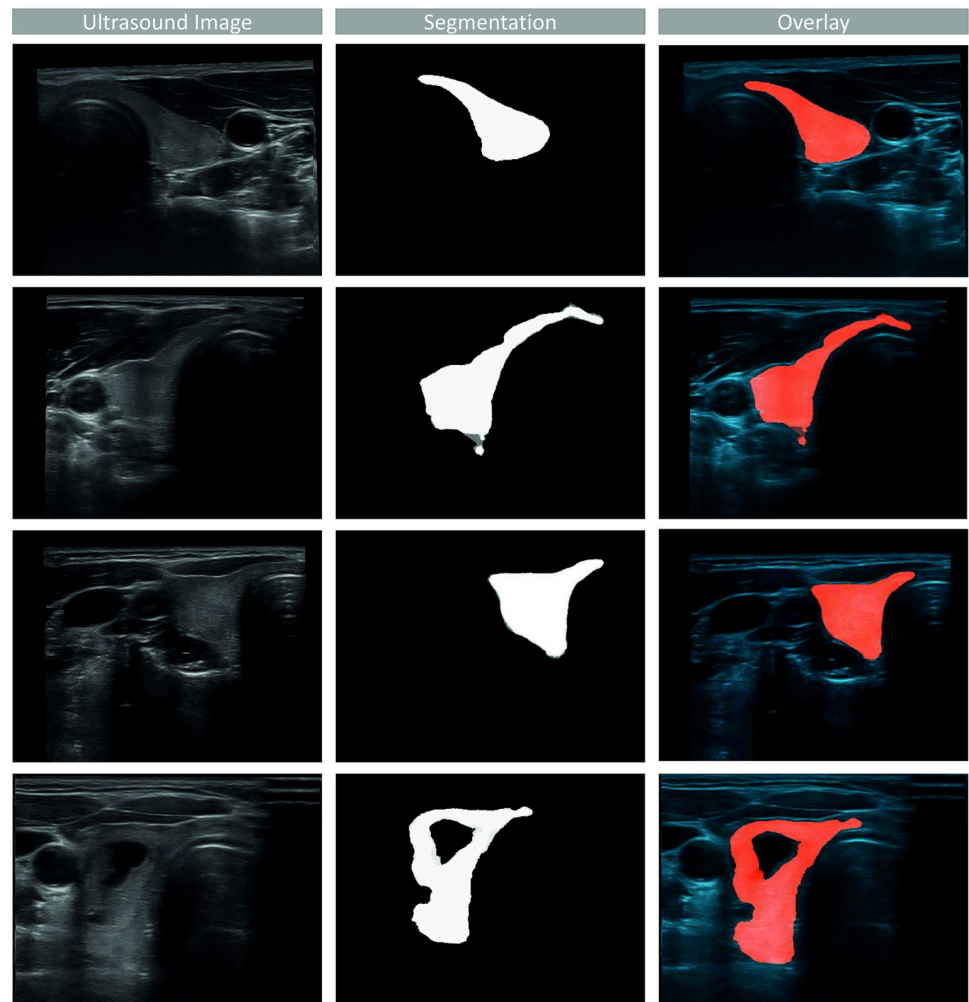


Fig 5. Ultrasound image (left), segmentation (middle) and overlay (right) of a good segmentation (first row), a segmentation which segmented too many voxels of the isthmus (second row), a segmentation with too few voxels of the isthmus segmented (third row) and a segmentation with a segmented nodule (fourth row).

<https://doi.org/10.1371/journal.pone.0268550.g005>

other with only unseen data (13 volunteers) (Group 2) to be able to detect a potential bias. To estimate the intraobserver variability of the acquisitions, we used a formula from Lyshchik et al. [5]. A t-test was performed with 2D, and 3D mean values to decide on whether both samples are significantly different. Additionally, we estimated the intraobserver variability as originally proposed by Choi et al. [18]. We examined the differences between two volumes of the same measurement sequence on a Bland-Altman plot [27]. We show the mean and the 95 limit of agreement ($\pm 1.96 * SD$). The interobserver variability was estimated by applying the method proposed by Choi et al. [18]. Here, the differences between volume measurements for each combination of two physicians were plotted in a Bland-Altman plot (Fig 7). Further, we applied paired samples t-tests to evaluate the difference between the mean volume measurements for each physician combination of two. In our case the objective was to determine whether the mean difference between the different measurements of the same volunteer is zero. We chose a significance level of $\alpha = 0.05$ for all tests in both groups. To compare 2D and 3D US volume measurements to MRI, we applied paired samples t-tests between MRI the first

Table 1. Mean and standard deviation (mean±SD) in percent (%) of the intraobserver variability for each MD in 2D and 3D for both groups (Gr. 1 and 2). The last column shows whether a significant difference exists between the averaged 2D and 3D volume estimation sets, being the significant different p-values in **bold**.

MD	Exp. (a)	2D US		3D US		p-value	
		Gr. 1	Gr. 2	Gr. 1	Gr. 2	Gr. 1	Gr. 2
1	6	12 ± 5	14 ± 10	12 ± 10	11 ± 8	.811	.081
2	4	18 ± 10	13 ± 9	15 ± 10	24 ± 25	.003	.040
3	1	15 ± 10	19 ± 10	8 ± 8	15 ± 12	<.001	<.001

<https://doi.org/10.1371/journal.pone.0268550.t001>

scan sequence from each MD with the same significance level resulting in the same critical values per group.

Results

We will present all results for Group 1 and Group 2 separately.

Intraobserver variability

Intraobserver variability is summarized in [Table 1](#). [Fig 6](#) depicts the Bland-Altman plots for the intraobserver variability of each MD with the 2D data on the left and the 3D data on the right for Group 2. Significant differences were found comparing the sets from MD2 and MD3 in both Groups 1 and 2.

Interobserver variability

The interobserver variability can be seen in [Table 2](#) for both groups. The Bland-Altman plots regarding the interobserver variability for Group 2 are depicted in [Fig 7](#).

Volumetry compared to MRI data

For the US volume, the first scan series per MD was chosen and compared to the MRI volume. [Table 3](#) summarizes these comparisons. The mean MRI volume overall volunteers of Group 1 was 7.73ml (SD = 2.71), while for Group 2, it was 7.00ml (SD = 3.48).

Time

Time needed for the 2D scans was on average 58.8 sec (SD 11.05 sec) for MD1, 55.0 sec (SD 8.8 sec) for MD2 and for 35.7 sec (SD 5.8 sec) MD3, while the 3D scans were performed on average in 26.0 sec (SD 5.5 sec) by MD1, 21.1 sec (SD 5.9 sec) by MD2 and in 21.6 sec (SD 5.2 sec) by MD3.

Discussion

Comparison with state-of-the-art

Intra- and interobserver variability in thyroid volumetry is a fairly well analysed problem. Vulpoi et al. had three MDs examine 30 children with 2D US. One physician scanned 25 patients twice to gather data on intraobserver variability, yielding 6.29% (SD = 6.12%). The interobserver difference was 9.51% (SD = 8.8%) [16]. Lee et al. studied intraobserver and interobserver variability between doctors with different experience levels (10 years and 6 months) on 122 nodules in 73 volunteers. The two MDs had a similar intraobserver variability -11.4% (SD = 10.5%) and 11.3% (SD = 8.5%). The interobserver variability amounted to 15.3% (SD = 16.6%) [17]. In the study of Choi et al. the thyroid nodule volume of 73 patients with 85

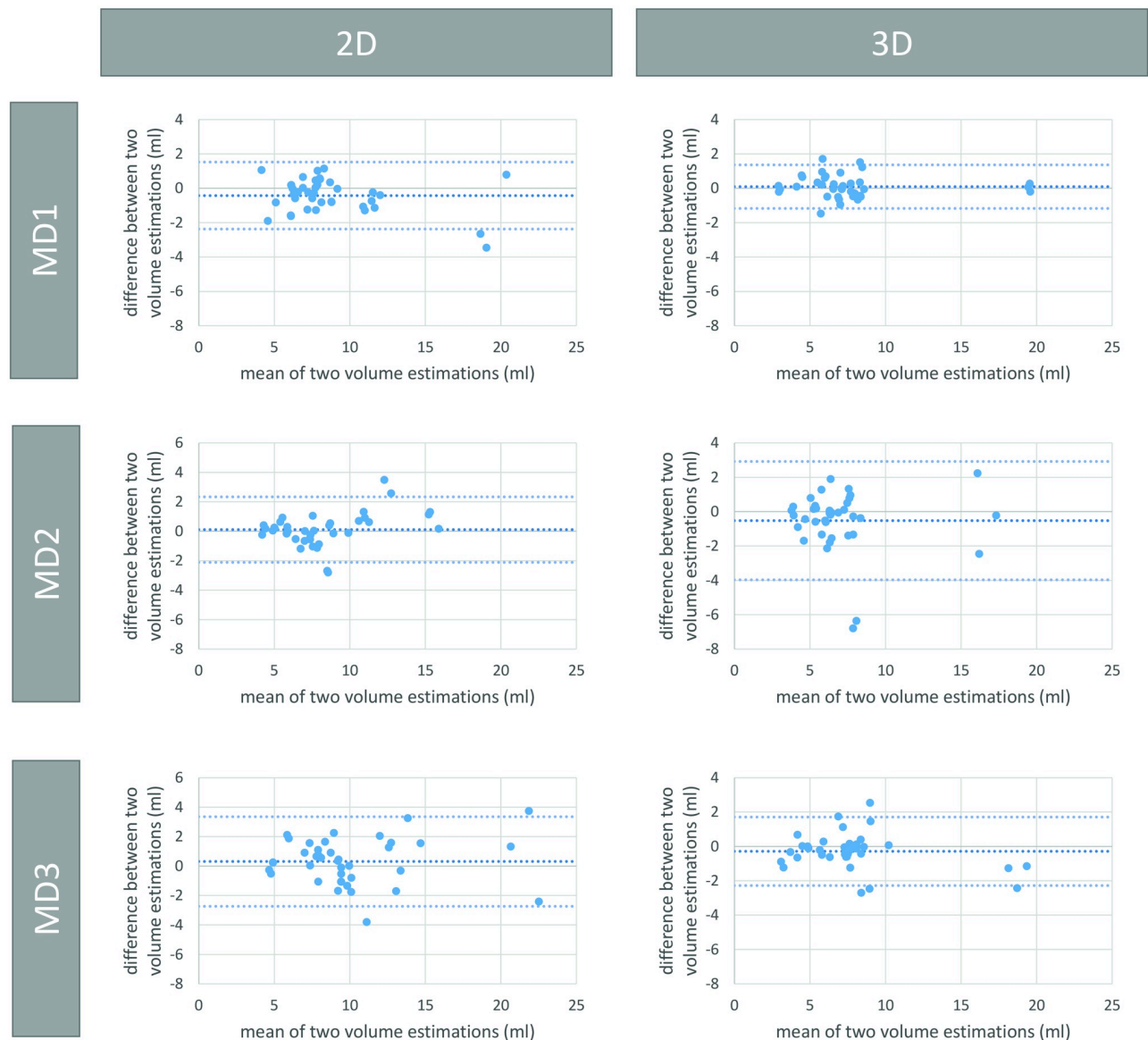


Fig 6. Bland-Altman plots for the intraobserver variability of each medical doctor from Group 2. All values are given in ml. Row 1, 2 and 3 show results for MD 1, 2, and 3, respectively. The columns show the results for 2D US (left) and 3D US measurements (right).

<https://doi.org/10.1371/journal.pone.0268550.g006>

nodules was measured by two MDs with different levels of experience twice. The thyroid nodules were divided into two groups: nodules with a diameter smaller than 2cm and those with a diameter bigger than 2cm. In the first group the interobserver variability was 7.0% ($SD = 4.9\%$) and -5.1% ($SD = 3.6\%$) in the second group [17, 18].

Our study focused on the total thyroid volume. In comparison to the results reported by Vulpoi et al., Lee et al. and Choi et al., our results for intra- and interobserver variability in 2D US (MD1 14%, MD2 13%, MD3 19% and MD1-MD2 6.5%, MD1-MD3 -13.12% , MD2-MD3 -19.31%) are slightly higher.

Lyshchik et al. conducted two studies on comparing the volume of thyroid nodules in children both with 2D and 3D US. In the first study one physician scanned 102 children with 129

Table 2. Mean and standard deviation (mean±SD) in milliliters (ml) of the interobserver variability of two MDs in 2D and 3D for both groups. Significant different p-values are marked in bold.

MDs	2D US		p-value	
	Gr. 1	Gr. 2	Gr. 1	Gr. 2
1/2	1.00 ± 1.35	0.58 ± 1.53	.005	.140
1/3	-2.49 ± 1.62	-1.33 ± 1.59	<.001	.002
2/3	-3.49 ± 1.43	-1.89 ± 2.04	<.001	.002
MDs	3D US		p-value	
1/2	0.16 ± 1.40	0.52 ± 1.62	.582	.176
1/3	-0.54 ± 1.03	-0.17 ± 1.03	.017	.722
2/3	-0.70 ± 1.15	-0.70 ± 1.68	.007	.057

<https://doi.org/10.1371/journal.pone.0268550.t002>

nodules three times, and delineated their thyroids. The delineated thyroids were controlled by a second expert. The second study included 47 children prior to thyroidectomy. Here, the thyroid weights after surgery were used as ground truth. The nodule volumes were calculated with the ellipsoid formula for 2D and multiplanar volume approximation, respectively manual planimetry, for the 3D scans. Both studies showed very similar results. 3D US showed a lower user dependency than 2D US, had a more accurate thyroid nodule volumetry, was less dependent on nodule size and performed better on irregular nodule outlines. In the first study the intraobserver variability was 16.1% ($SD = 0.7\%$) for 2D US and 5.9% ($SD = 0.3\%$) for 3D US. The accuracy for 2D US was 15.9% and 2.8% for 3D US. In the second study the intraobserver variability amounted to 14.4% ($SD = 1.9\%$) in 2D US and to 3.36% ($SD = 0.25\%$) for 3D US. The accuracy amounted to 15.3% for 2D US and 5.2% for 3D US. The second study showed no significant variation between the thyroid weight and both US modalities [5, 20]. One limitation of both their studies is that only one physician acquired the data. In our study intraobserver variability for 2D US yielded similar results to Lyshchik et al. Our results for 3D US however are worse. We speculate that the reason for this could be either (a) our network performance in segmenting the thyroid lobes, (b) breathing or movement distortions as we are using tracked ultrasound instead of a two-dimensional array ultrasound or (c) tracking errors.

With respect to MRI, Reinartz et al. compared 2D US thyroid volumetry to MRI (T1 FFE). Three different experienced MDs performed US scans on patients prior to RIT. The thyroid volumes were calculated by using the ellipsoid formula for the 2D US measurements and by manual segmentation. A significant difference of 22.7% (10.4ml), between the MRI-ROI and mean 2D US measurements was found. No significant interobserver variability among the MDs could be determined [15]. This finding is in contradiction with our results. In our case the MRI volumes were on average 21.3% smaller than the ones obtained from 2D US and 4.99% smaller than the ones obtained from 3D US. This may have to do with the fact that different MRI sequences, resolutions or sampling were used, but most likely by the criteria used for segmentation [23]. One study comparing 3D US to CT by Licht et al showed no relevant differences [28]. Rogers et al. analyzed length, diameters and volume measurements of four arteries from a pig using 3D tUS, B-Mode US, CT and MRI. As validation, water immersion technique was used. B-Mode US had the largest error in volume estimation, tUS the smallest. The mean error in volume estimation was $-0.54 \pm 0.62ml$ for B-Mode, $-0.06 \pm 0.09ml$ for tUS, $0.01 \pm 0.18ml$ for CT and $-0.20 \pm 0.32ml$ for MRI. These results align with our results in showing that tracked 3D US shows a smaller difference to MRI than 2D US. Furthermore, it shows the second largest mean volume error in volume estimation for MRI. This aligns with our observation of a probable continuous underestimation of our MRI segmentations.

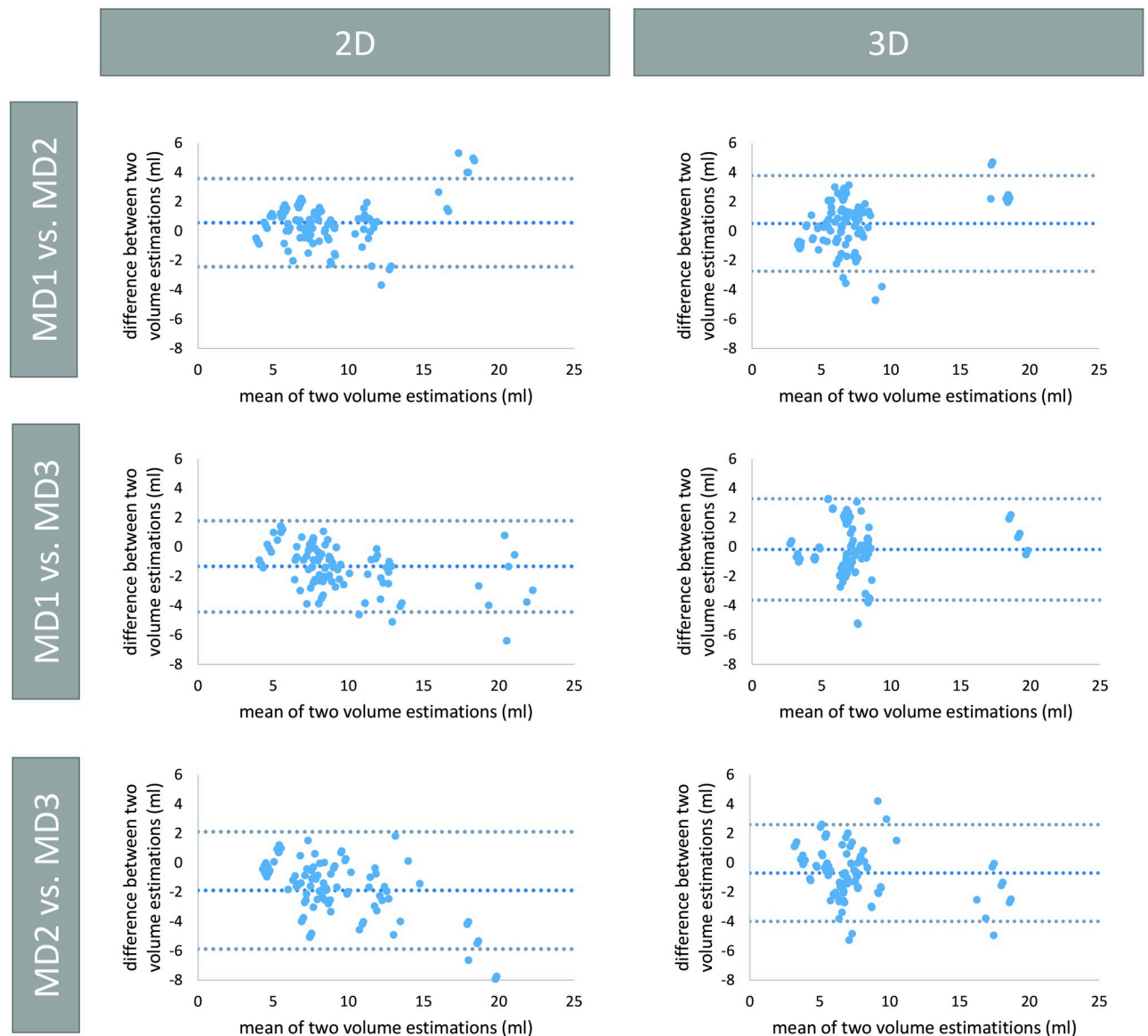


Fig 7. Bland-Altman plots for the interobserver variability between sets of two medical doctors from Group 2. All values are given in ml. Row 1, 2 and 3 show results between MD 1 and MD 2, MD 1 and MD 3, and MD 2 and MD 3, respectively. The columns show the results for 2D (left) and 3D measurements (right).

<https://doi.org/10.1371/journal.pone.0268550.g007>

Method evaluation

The current network was trained and evaluated on ultrasound images from one ultrasound machine, with slightly different imaging parameters (such as gain, and focus). Based on this data, the network can handle small changes in the ultrasound image. For more prominent changes in the imaging parameters (such as frequency, ultrasound transducer, speckle filtering, among others), resulting in more considerable changes in the ultrasound image, we believe that the CNN would have to be adjusted in its architecture and fine-tuned to increase its generalizability. Compared to the current clinical practice, the improved performance can be based on the nature of the 3D acquisition and the automatic segmentation. 3D US itself

Table 3. Mean and standard deviation (mean±SD) of the volume of the first scan series for each MD in 2D and 3D in milliliters (ml). Statistical significance with respect to the MRI volume is reported, being statistical significant p-values in **bold**.

MD	2D US		p-value	
	Gr. 1	Gr. 2	Gr. 1	Gr. 2
1	8.53 ± 2.40	8.38 ± 3.26	.048	.002
2	7.66 ± 2.12	8.19 ± 3.52	.880	.009
3	11.80 ± 2.41	10.09 ± 4.32	<.001	<.001
MD	3D US		p-value	
1	9.09 ± 1.80	7.62 ± 3.88	.018	.292
2	8.82 ± 2.35	6.70 ± 3.43	.029	.686
3	9.41 ± 2.26	7.55 ± 3.57	.003	.091

<https://doi.org/10.1371/journal.pone.0268550.t003>

makes a difference compared to 2D US as it removes the variability of the probe placement with respect to the anatomy through volume compounding. The automatic segmentation allows for a complete thyroid segmentation, which provides a shape closer to the actual thyroid form than an ellipsoid approximation.

Limitations of this study

Only young and healthy volunteers were enrolled, simplifying the segmentation and volumetry tasks. The CNN was trained with 16 volunteers only, allowing for a good Dice score but leaving space for improvements. These can include cross-validation as well as training on 3D input data instead of 2D images. The latter one might allow for additional spatial information which could improve the network performance; yet, it would require a significantly larger dataset. Visual analysis also showed non-optimal segmentation results in few scans (Fig 5). The network seems to be sensitive to the echogenicity of the US images, resulting in additional areas outside of the thyroid to be segmented and areas inside the thyroid to be not segmented. Furthermore, the network was trained on single lobes. This can introduce a potential deviation in the volume due to overlapping or non-segmented parts of the isthmus. It also has to be noted that not all MDs have used the 3D system before. A longer training could therefore increase the scanning quality.

Conclusion

Our study shows that 3D US outperforms 2D US significantly. This was consequently seen in both groups 1 and 2 denoting “training” and “new” data. Regarding the intraobserver variability, 3D US outperforms 2D US with 2 out of 3 MDs. In the interobserver variability 3D US increases the similarity of acquisition scans between two MDs in 2 out of 3 cases. The results also show that the improvement with 3D US is the biggest for the least experienced MD. We show that the 3D US volumetry is more accurate than 2D US using MRI as ground truth. The tracked 3D US acquisition is significantly faster as well. Therefore, the combination of tracked 3D US and automatic segmentation is not only feasible but offers a more accurate alternative to conventional US imaging. The labelled thyroid 3D US dataset will be freely available helping further projects in deep learning for thyroid volumetry within the research community. We suggest conducting similar studies on patients with thyroid morbidities to stratify the clinical benefit of our framework.

Acknowledgments

The Nuclear Medicine department at KRI thanks Piur Imaging GmbH (www.piurimaging.com; piur imaging GmbH, Hamburgerstr. 11/7, 1050 Vienna, Austria) for a free loan of a PIUR tUS system.

Author Contributions

Conceptualization: Markus Krönke, Desislava Dimova, James Nagarajah, Thomas Wendler.

Data curation: Markus Krönke, Desislava Dimova, Melanie Köhler, Gabriel Buschner, Lilit Schweiger, LEMONIA Konstantinidou.

Formal analysis: Markus Krönke, Christine Eilers, Desislava Dimova.

Methodology: Markus Krönke, Christine Eilers, Desislava Dimova, James Nagarajah, Thomas Wendler.

Project administration: Thomas Wendler.

Resources: Markus Krönke.

Software: Desislava Dimova, Melanie Köhler, LEMONIA Konstantinidou.

Supervision: Marcus Makowski, Nassir Navab, Wolfgang Weber, Thomas Wendler.

Validation: Desislava Dimova, LEMONIA Konstantinidou.

Visualization: Christine Eilers.

Writing – original draft: Markus Krönke, Christine Eilers, Desislava Dimova, Melanie Köhler, Gabriel Buschner, Lilit Schweiger, LEMONIA Konstantinidou, Thomas Wendler.

Writing – review & editing: Markus Krönke, Christine Eilers, Thomas Wendler.

References

1. Szumowski P., Mojsak M., Abdelrazek S., Sykała M., Amelian-Fitonowicz A., Jurgilewicz D. et al. Calculation of therapeutic activity of radioiodine in Graves' disease by means of Marinelli's formula, using technetium (99mTc) scintigraphy. *Endocrine*. 54, 751–756 (2016) <https://doi.org/10.1007/s12020-016-1074-7> PMID: 27553050
2. Brunn J., Block U., Ruf G., Bos I., Kunze W. & Scriba P. [Volumetric analysis of thyroid lobes by real-time ultrasound (author's transl)]. *Dtsch Med Wochenschr*. 106, 1338–1340 (1981) <https://doi.org/10.1055/s-2008-1070506> PMID: 7274082
3. Dietlein M., Grünwald F., Schmidt M., Schneider P., Verburg F. & Luster M. Radioiodtherapie bei benignen Schilddrüsenerkrankungen (Version 5)*: DGN-Handlungsempfehlung (S1-Leitlinie). *Nuklearmedizin*. 55, 213–220 (2016), <http://www.thieme-connect.de/DOI/DOI?10.3413/Nukmed-0823-16-04> PMID: 27922153
4. Mariani G., Tonacchera M., Grosso M., Orsolini F., Vitti P. & Strauss H. The Role of Nuclear Medicine in the Clinical Management of Benign Thyroid Disorders, Part 1: Hyperthyroidism. *J Nucl Med*. 62, 304–312 (2021) <https://doi.org/10.2967/jnumed.120.243170> PMID: 33008929
5. Lyschchik A., Drozd V., Schloegl S. & Reiners C. Three-dimensional ultrasonography for volume measurement of thyroid nodules in children. *J Ultrasound Med*. 23, 247–254 (2004) <https://doi.org/10.7863/jum.2004.23.2.247> PMID: 14992362
6. Kojcev R., Khakzar A., Fuerst B., Zettinig O., Fahkry C., DeJong R., et al. On the reproducibility of expert-operated and robotic ultrasound acquisitions. *Int J Comput Assist Radiol Surg*. 12, 1003–1011 (2017) <https://doi.org/10.1007/s11548-017-1561-1> PMID: 28321804
7. Andermann P., Schlögl S., Mäder U., Luster M., Lassmann M. & Reiners C. Intra- and interobserver variability of thyroid volume measurements in healthy adults by 2D versus 3D ultrasound. *Nuklearmedizin*. 46, 1–7 (2007) <https://doi.org/10.1055/s-0037-1616621> PMID: 17299648

8. Chang C., Lei Y., Tseng C. & Shih S. Thyroid Segmentation and Volume Estimation in Ultrasound Images. *IEEE Transactions On Biomedical Engineering*. 57, 1348–1357 (2010) <https://doi.org/10.1109/TBME.2010.2041003> PMID: 20172782
9. Poudel P., Illanes A., Sheet D. & Friebe M. Evaluation of Commonly Used Algorithms for Thyroid Ultrasound Images Segmentation and Improvement Using Machine Learning Approaches. *Journal Of Healthcare Engineering*. (2018) <https://doi.org/10.1155/2018/8087624> PMID: 30344990
10. Illanes Parametrical modelling for texture characterization—A novel approach applied to ultrasound thyroid segmentation. *PLOS ONE*. 14, 1–17 (2019), <https://doi.org/10.1371/journal.pone.0211215> PMID: 30695052
11. China D., Illanes A., Poudel P., Friebe M., Mitra P. & Sheet D. Anatomical Structure Segmentation in Ultrasound Volumes Using Cross Frame Belief Propagating Iterative Random Walks. *IEEE Journal Of Biomedical And Health Informatics*. 23, 1110–1118 (2019) <https://doi.org/10.1109/JBHI.2018.2864896> PMID: 30113902
12. Mahmoodian, N., Poudel, P., Illanes, A. & Friebe, M. Higher Order Statistical Analysis for Thyroid Texture Classification and Segmentation in 2D ultrasound Images. *2019 41st Annual International Conference Of The IEEE Engineering In Medicine And Biology Society (EMBC)*. pp. 5832–5835 (2019)
13. Poudel, P., Illanes, A., Sadeghi, M. & Friebe, M. Patch Based Texture Classification of Thyroid Ultrasound Images using Convolutional Neural Network. *2019 41st Annual International Conference Of The IEEE Engineering In Medicine And Biology Society (EMBC)*. pp. 5828–5831 (2019)
14. Webb J., Meixner D., Adusei S., Polley E., Fatemi M. & Alizad A. Automatic Deep Learning Semantic Segmentation of Ultrasound Thyroid Cineclips Using Recurrent Fully Convolutional Networks. *IEEE Access*. 9 pp. 5119–5127 (2021) <https://doi.org/10.1109/access.2020.3045906> PMID: 33747681
15. Reinartz P., Sabri O., Zimny M., Nowak B., Cremerius U., Setani K. et al. Thyroid volume measurement in patients prior to radioiodine therapy: comparison between three-dimensional magnetic resonance imaging and ultrasonography. *Thyroid*. 12, 713–717 (2002) <https://doi.org/10.1089/105072502760258695> PMID: 12225640
16. Vulpoi C., Mogoş V., Ciobanu D., Preda C., Ungureanu M., Cristea C., et al. [Thyroid ultrasonography. Interobserver and intraobserver variations]. *Rev Med Chir Soc Med Nat Iasi*. 111, 864–871 (2007) PMID: 18389771
17. Lee H., Yoon D., Seo Y., Kim J., Baek S., Lim K., et al. Intraobserver and Interobserver Variability in Ultrasound Measurements of Thyroid Nodules. *J Ultrasound Med*. 37, 173–178 (2018) <https://doi.org/10.1002/jum.14316> PMID: 28736947
18. Choi Y., Baek J., Hong M. & Lee J. Inter-observer variation in ultrasound measurement of the volume and diameter of thyroid nodules. *Korean J Radiol*. 16, 560–565 (2015) <https://doi.org/10.3348/kjr.2015.16.3.560> PMID: 25995685
19. Schlögl S., Andermann P., Luster M., Reiners C. & Lassmann M. A novel thyroid phantom for ultrasound volumetry: determination of intraobserver and interobserver variability. *Thyroid*. 16, 41–46 (2006) <https://doi.org/10.1089/thy.2006.16.41> PMID: 16487012
20. Lyschchik A., Drozd V. & Reiners C. Accuracy of three-dimensional ultrasound for thyroid volume measurement in children and adolescents. *Thyroid*. 14, 113–120 (2004) <https://doi.org/10.1089/105072504322880346> PMID: 15068625
21. Malago R., D'Onofrio M., Ferdeghini M., Mantovani W., Colato C., Brazzarola P., et al. Thyroid volumetric quantification: comparative evaluation between conventional and volumetric ultrasonography. *J Ultrasound Med*. 27, 1727–1733 (2008) <https://doi.org/10.7863/jum.2008.27.12.1727> PMID: 19022998
22. Schlögl S., Werner E., Lassmann M., Terekhova J., Muffert S., Seybold S. et al. The use of three-dimensional ultrasound for thyroid volumetry. *Thyroid*. 11, 569–574 (2001) <https://doi.org/10.1089/105072501750302877> PMID: 11442004
23. Freesmeyer M., Wiegand S., Schierz J., Winkens T. & Licht K. Multimodal evaluation of 2-D and 3-D ultrasound, computed tomography and magnetic resonance imaging in measurements of the thyroid volume using universally applicable cross-sectional imaging software: a phantom study. *Ultrasound Med Biol*. 40, 1453–1462 (2014) <https://doi.org/10.1016/j.ultrasmedbio.2014.02.013> PMID: 24768486
24. Guha Roy A., Conjeti S., Navab N., Wachinger C. & Alzheimer's Disease Neuroimaging Initiative QuickNAT: A fully convolutional network for quick and accurate segmentation of neuroanatomy. *Neuroimage*. 186 pp. 713–727 (2019) <https://doi.org/10.1016/j.neuroimage.2018.11.042> PMID: 30502445
25. Ronneberger, O., Fischer, P. & Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. *Medical Image Computing And Computer-Assisted Intervention—MICCAI 2015*. pp. 234–241 (2015)
26. Castro, E., Cardoso, J. & Pereira, J. Elastic deformations for data augmentation in breast cancer mass detection. *2018 IEEE EMBS International Conference On Biomedical Health Informatics (BHI)*. pp. 230–234 (2018)

27. Bland J. & Altman D. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*. 1, 307–310 (1986) [https://doi.org/10.1016/S0140-6736\(86\)90837-8](https://doi.org/10.1016/S0140-6736(86)90837-8) PMID: [2868172](https://pubmed.ncbi.nlm.nih.gov/2868172/)
28. Licht K., Darr A., Opfermann T., Winkens T. & Freesmeyer M. 3D ultrasonography is as accurate as low-dose CT in thyroid volumetry. *Nuklearmedizin*. 53, 99–104 (2014) <https://doi.org/10.3413/Nukmed-0615-13-08> PMID: [24276677](https://pubmed.ncbi.nlm.nih.gov/24276677/)