A generalized approach for automatic 3-D geometry assessment of blood vessels in transverse ultrasound images using convolutional neural networks

A generalized approach for automatic 3-D geometry assessment of blood vessels in transverse ultrasound images using convolutional neural networks

Joerik de Ruijter, Judith J.M. Muijsers, Frans N. van de Vosse, Marc R.H.M. van Sambeek and Richard G.P. Lopata

***Abstract*—Accurate 3-D geometries of arteries and veins are important clinical data for diagnosis of arterial disease and intervention planning. Automatic segmentation of vessels in the transverse view suffers from the low lateral resolution and contrast. Convolutional neural networks are a promising tool for automatic segmentation of medical images, outperforming the traditional segmentation methods with high robustness. In this study, we aim to create a general, robust, and accurate method to segment the lumen-wall boundary of healthy central and peripheral vessels in large field-of-view freehand ultrasound (US) datasets. Data were acquired using freehand US, in combination with a probe tracker. A total of $\pm$ 36000 cross-sectional images, acquired in the common, internal, and external carotid artery (N = 37), in the radial, ulnar artery, and cephalic vein (N = 12), and in the femoral artery (N = 5) were included. To create masks (of the lumen) for training data, a conventional automatic segmentation method was used. The neural networks were trained on a) data of all vessels and b) the carotid artery only. The performance was compared and tested using an open access dataset. The Recall, Precision, DICE, and the intersect-over-union (IoU) were calculated. Overall, segmentation was successful in the carotid and peripheral arteries. The Multires U-net architecture performs best overall with DICE = 0.93 when trained on the total dataset. Future studies will focus on the inclusion of vascular pathologies.**

***Index Terms*—Convolutional Neural Network, Machine Learning, Medical Image Segmentation, Vascular Ultrasound.**

I. Introduction

Accurate 3-D geometries of arteries are important clinical data for grading stenosis, assessing aneurysmatic disease, and intervention planning. For instance, the visualisation of the vasculature can contribute to arteriovenous fistula placement for hemodialysis, or stent placement in lower extremity atherosclerotic lesions. Ultrasound (US) is the modality of choice, due to its non-invasive nature, ease-of-use, and low cost. However, automatic segmentation suffers from the low lateral resolution and contrast. Deep convolutional neural networks (CNNs) are a promising tool for semantic segmentation of medical images. Long, et al. [1] first presented a fully convolutional network (FCN) for semantic segmentation, outperforming the traditional segmentation methods and showing a high level of robustness. The U-net was designed for medical image segmentation, and was proven to be successful in many other applications [2]. U-nets perform well even without huge amounts of labelled training data and are now the preferred approach for medical image segmentation [3].

Since the introduction of the U-net, many variations were introduced to optimize the performance of this type of neural network. The U-net combines low level detail information with high level semantic information, by the connections between decoding and encoding blocks. However, to train a network, large annotated datasets are required, which are typically difficult to obtain, especially in biomedical imaging or when developing a novel, not yet clinically used method. A commonly used solution to increase the amount of data is data augmentation, here the existing data is modified using different image operations.

In the field of vascular ultrasound (US), convolutional neural networks are already under investigation for different purposes. Smistad, et al. [4] used a CNN to classify vessels detected by their Vessel Candidate Search algorithm. Jain, et al. [5] used a faster Region-CNN to localize the common carotid artery in transverse B-mode ultrasound images. Several groups have used a CNN to measure the intima - media thickness of the common carotid artery in longitudinal images [6], whereas [7] used a similar approach aimed at characterizing the composition of carotid plaques. Zhou, et al. [8] used a Residual U-net to segment the lumen-intima boundary in 2-D carotid images from a single volume 3-D dataset. Xie, et al. [9] used a dual-path encoder U-net to segment the carotid lumen in longitudinal B-mode images. Smistad, et al. [10] used a U-net to detect blood vessels and nerves in transverse images during ultrasound-guided axillary nerve block procedures. Mishra, et al. [11] deployed a deeply supervised FCNN to show vessel segmentation in liver ultrasound images and IVUS images of coronary arteries.

In this study, we aim to create a generalized, robust, and accurate method to segment the lumen-wall boundary for larger segments of central and peripheral arteries in large field-of-view datasets using CNNs. The goal is to investigate the influences of addition of various vessels to the training data versus an augmentation strategy on the performance on a single artery, in this case the CCA. A large dataset was created

This study was funded by Stichting Lijf and Leven. Computational work was carried out on the Dutch national e-infrastructure, supported by the SURF Cooperative.

Joerik de Ruijter, Department of Biomedical Engineering, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands (e-mail: j.d.ruijter@tue.nl).

Marc R.H.M. van Sambeek, department of Surgery, Catharina Hospital Eindhoven, P.O. box 1350, 5602 ZA Eindhoven, the Netherlands (e-mail: marc.v.sambeek@catharinaziekenhuis.nl).

Richard G.P. Lopata, Department of Biomedical Engineering, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands (e-mail: r.lopata@tue.nl).

from stacks of transverse US images of different arteries and veins: 1) the carotid artery including the internal carotid artery (ICA), the external carotid artery (ECA), and the common carotid artery (CCA); 2) the femoral artery (FA) in the upper leg; and 3) the vasculature in the lower arm, including the radial artery (RA), ulnar artery (UA), and the cephalic vein (CV). The acquisitions were performed by making a sweep along the subject's artery or vein, this maximizes the number of images of each subject, whilst still introducing variations in vessel size, vessel location, surrounding tissue, contrast etc. and allows for an assessment of 3-D geometry for a large field-of-view.

To reduce manual labor, the masks were created using an existing semi-automatic segmentation method [12]. To ensure acceptable quality of masks, erroneous masks were manually removed. This does not always result in pixel-perfect masks, however in (vascular) ultrasound imaging this is also hard to achieve with manual annotation since there is high inter- and intra-observer variability [13].

The neural networks were trained with two sets of training data: carotid data only, and data originating from all arteries. Besides the Standard U-net, four modifications of the U-net were tested: Residual U-net [14], Compact U-net [15], Dense U-net [15], and Multires U-net [16]. The performance of the different network architectures was compared and tested using data from an online carotid database [17] and manually annotated in-house datasets. Finally, the best performing neural network was used to segment blood vessels in large field-of-view, probe-tracked, 3-D datasets to obtain a large part of their full, 3-D geometry.

II. Materials and Methods

A. Data acquisition

Training data were acquired using freehand US. The arteries were primarily imaged in the transverse view, while the probe was moved in the out-of-plane direction. This results in obtaining multiple images of each subject with different image features. For carotid and lower extremity imaging a MylabOne ultrasound system (ESAOTE EUROPE, Maastricht, NL) with a linear array ($f_c = 7.5$ MHz) was used. For imaging the upper extremities a Mylab70 ultrasound system (ESAOTE EUROPE, Maastricht, NL) with a linear array ($f_c = 10$ MHz) was used, since the arteries are smaller and a higher resolution (and thus transmit frequency) was needed. All probes were connected to a magnetic probe tracking device (CUREFAB CS, Münich, Germany) for 3-D reconstructions. Two magnetic field sensors were mounted on the ultrasound probe, which were able to sense the magnetic field generated by the magnetic field emitter. Utilizing the information of the magnetic field emitters, the position and rotations (6 Degrees-of-Freedom) of the transducer relative to the magnetic field emitter are calculated. The position of the 2 sensors relative to the probe head (and thus the US image) were calibrated using a pyramid phantom. For all applications, a slow sweep at a speed of $v =\sim 5mm/s$ was performed along the subjects artery, so that the vessel was continuously (fully) visible in a transverse view. The left and right carotid artery were both imaged, starting distal to the bifurcation, moving to the proximal part while the subject was in the supine position. To measure the vasculature in the forearm, the volunteer was asked to place their non-dominant arm in supination on a handcrafted holder. The radial artery, ulnar artery, and cephalic vein were measured from proximal to distal position in separate sweeps, following each vessel. To measure the femoral artery, the volunteer was laying down on a bed. Data acquisition was performed in volunteers, see Table 1. All subjects gave their informed consent. All separate studies were approved by the local ethics committee of the Catharina Ziekenhuis Eindhoven (MEC-U, Eindhoven, NL). The B-mode cineloops, with a frame rate of 25 Hz, were extracted from the probe tracking device, including the corresponding spatial coordinates of the probe during the sweep.

TABLE I
AN OVERVIEW OF NUMBER OF SUBJECTS, ACQUISITIONS AND THE TOTAL NUMBER OF TRAINING IMAGES.

Type	volunteers	acquisitions	images
Carotid	37	67	18379
Upper extremities	12	36	16069
Lower extremities	5	10	1705

B. Training data preparation

An (semi) automatic segmentation approach was used to find the lumen-wall boundaries in the training data. Since all images were acquired using a slow sweep in combination with a probe tracking device, frame by frame segmentation was performed, using information from the previous frame to predict the location and size of the artery of interest in the next frame. This automatic segmentation algorithm was recently published [12]. In short, the algorithm uses the Star-Kalman method to approximate the center and the size of the vessel(s) for every frame [18]. Images were filtered with a Gaussian low-pass filter before conversion into the 2-D monogenic signals. Multi-scale asymmetry features were extracted from these data, enhancing low lateral wall-lumen contrast. These images, in combination with the initial ellipse contours, were used for an active deformable contour model to segment the lumen-wall boundary [19]. The output of the algorithm are contours (100 points) for every artery. A mask was generated (intensity = 1 for lumen pixels, intensity = 0 for background pixels). Every contour was checked visually; if the algorithm failed to track the artery correctly, the erroneous images/masks were removed from the total dataset for training. The acquired US images had different dimensions and aspect ratio's, due to the use of different US devices and different depth settings. However, all input data are required to have the same size. In general, all images had more pixels in depth compared to the width. Hence, all images were cropped, i.e., the bottom was removed so that every image is square. Next, the images and masks were resized to 128 by 128 pixels. This approach was not feasible for upper/lower extremity images, since the artery could be located at the bottom of the image. Here, the full image was resized, resulting in non-equidistant pixel spacing. Padding was also considered, however the padded area does

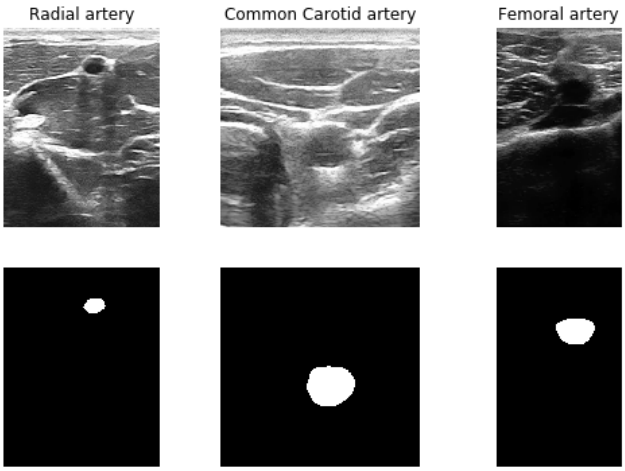

Fig. 1. Examples of B-mode image and training mask for the radial, common carotid, and femoral artery.

not contribute to the pixel classification, and would require an extra step for the learning algorithm. Figure 1 shows three examples of the training data.

C. Network architectures

The U-net and modifications of this architecture have shown to be very suitable for semantic segmentation. Figure 2 shows a schematic overview of the U-net architecture, including the encoding and decoding blocks. Besides the Standard U-net, different state-of-the-art modifications have been tested: Residual U-net [14], Compact & Dense U-net [15], and Multires U-net [16]. Research showed that deeper networks, networks with more convolutional blocks, perform better. However, deeper networks are more difficult to train, due to vanishing gradients. The residual networks have the strengths of skipping connections to avoid this problem. In the Residual U-net, the Standard U-net block has been replaced by a Residual block, see Figure 3. Densely connected networks have interconnections after each convolution. The Compact version has only one connection in each building block. In the standard building block, a 3 x 3 convolution operation is used twice, which resembles a 5 x 5 convolution layer. In the Multires building block, next to the 5 x 5 convolution layer, a 3 x 3 and a 7 x 7 convolution layer were added. This enables the network to capture image features of different dimensions. Figure 3 shows the building blocks for every archetype. In case of encoding, the last operation is so-called max pooling, the input is downsampled by taking the maximum value of a 2 by 2 window. For the upsampling scheme a transposed 2-D convolution is used, also known as deconvolution. For all U-net archetypes 32 convolutional kernels were used in the first layer (see Figure 2). The total number of parameters of each architecture is shown in Table II. The implementation of the architectures was adapted from [20].

D. Training

Each network was trained using two groups of training data: data of the carotid artery and data of all arteries. Dataset

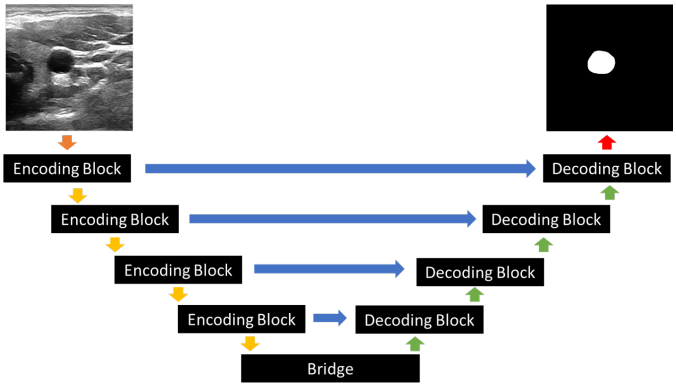

Fig. 2. An overview of the U-net architecture. The encoding and decoding blocks are shown in Figure 3.

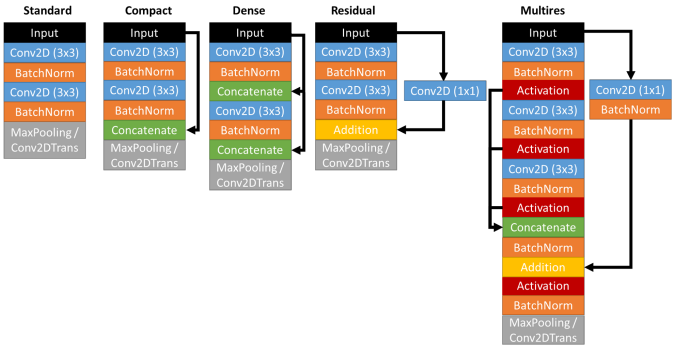

Fig. 3. Building blocks for the different U-net archetypes used in this study.

"Carotid" consists of ± 18k images, dataset "All" of ± 36k images. To make a fair comparison, dataset "Carotid" was doubled to 36k images using data augmentation. Due to the sweep acquisition, consecutive frames can be similar, to introduce more variety in the data, augmentation was also applied to all the original data. Six augmentation techniques were used: Deformation, rotation, zoom, gamma correction, addition of Gaussian noise, and contrast adjustments. The data augmentation was implemented using the Python library 'batchgenerators' [21]. The parameters are summarized in Table III.

The neural networks were trained in a k-folds cross-validation approach. The training set was randomly divided in 5 mutually exclusive subsets of equal size. Each network was trained 5 times, subsequently with a different subset as validation set. The networks were trained for 100 epochs after which the validation loss improvements were found to be minor. The batch-size was set to 10 and learning rate to

TABLE II
THE TOTAL NUMBER OF PARAMETERS FOR EACH ARCHITECTURE USED IN THIS STUDY.

Architecture	Total parameters
Standard	7.8M
Compact	10.7M
Dense	15.8M
Residual	8.1M
Multires	7.3M

TABLE III
THE PARAMETERS USED FOR DATA AUGMENTATION.

Augmentation type	Parameters
Rotation (range in degrees)	$[-15°, 15°]$
Deformation (α, σ)	$\alpha = [0, 200]$ $\sigma = [9, 13]$
Zoom	$[0.75, 1.25]$
Gaussian noise	$[0.0, 0.05]$
Contrast	$[0.3, 3]$
Probability	0.2

$1e^{-5}$. The $DICE_{loss}$ was used as loss metric (Equation 3), the weights were saved when the validation loss improved after an epoch. Training was performed on a GPU (NVIDIA Tesla K40m) provided by the SURF Cooperative (Amsterdam, The Netherlands) .

E. Metrics

In the training sets there is an imbalance in classes (number of background pixels $\gg$ number object pixels). The DICE index (DICE) (also known as the Similarity Index) is a metric that is insensitive to this imbalance problem. An other commonly used quality metric is intersection over union (IoU). The equations for the DICE and IoU are given by:

$$DICE = \frac{2 \cdot A \cap B}{A + B} \tag{1}$$

$$IoU = \frac{A \cap B}{A \cup B} \tag{2}$$

Here, A and B are the set of pixels classified as 'vessel' in the ground truth and prediction, respectively. The loss function used for training and validation is given by:

$$DICE_{loss} = 1 - DICE \tag{3}$$

Besides the pixel based metrics, the network's ability to detect the vessel is of interest. Recall (R) describes completeness of the model's positive predictions versus the ground truth. Precision (P) is a metric that shows the correctness of the models positive predictions versus the ground truth. A predicted object is defined as a group of more than 10 connected pixels in the binary prediction image. A prediction is considered true positive (TP) when the DICE between the predicted object and the corresponding ground truth is higher than 0.5, otherwise the predicted object is counted as false positive (FP). When a ground truth object does not have a corresponding prediction with a DICE higher than 0.5, this is considered as false negative (FN).

$$R = \frac{N_{TP}}{N_{TP} + N_{FN}} \tag{4}$$

$$P = \frac{N_{TP}}{N_{TP} + N_{FP}} \tag{5}$$

here, N_{TP} is the total number of true positive predictions and N_{FP} is the total number of false positive predictions. N_{FN} is the number of false negative predictions [22]. Test set 1 is an online CCA database that only provides a detection box (center coordinates, box height and box width). These detection boxes, used as ground truth, are larger than the actual vessel, therefore the DICE threshold was reduced to 0.25 for Test Set 1.

TABLE IV
OVERVIEW OF THE TEST SETS USED IN THIS STUDY.

Name	Source	images	Annotation	Location
Test Set 1	Online	971	Location only	CCA
Test Set 2	Online	269	Manual	CCA
Test Set 3	In-house	94	Manual	CCA
Test Set 4	In-house	100	Manual	Arm

F. Test sets

The training data cannot be regarded as ground truth data, since the masks were generated using a semi-automatic method. Hence, these data cannot be used as test set. Four test sets were used to test the segmentation results of each trained network. Test Set 1 and 2 originate from an online CCA database [17], consisting of 971 transverse US images of the CCA acquired with two different US machines: an Ultrasonix (538 images) and Toshiba US system (433 images). The database included the position of the center and radius of the carotid artery for each cross-section. For each image it was tested if the neural network was able to detect the position of the artery. Test Set 2 contains 269 images obtained with an Ultrasonix US device, these images were segmented manually. Test Set 3 is a CCA dataset, acquired in-house with a MylabOne, which were also segmented manually. The manual annotations were compared with the predicted images, and the DICE, IoU, Recall, and Precision were calculated. Additionally, an in-house dataset consisting of 100 images of the cephalic vein, radial artery, and the ulnar artery of two volunteers, that were not included in the training data, were segmented manually (Test Set 4). Note that in this set only the vessel of interest was annotated. It was not feasible to annotate all small vessels that appear during the sweep, since it was not always clearly recognizable to the observer. A summary of the test sets is given in Table IV.

G. Geometry reconstruction

In this study, the resulting lumen geometry was reconstructed to visualize the vasculature in 3-D. Each sweep dataset, containing images and probe tracking coordinates for each image, is exported to MATLAB 2019a (MathWorks, USA). The segmentation is performed by the Compact U-net. An empty voxel volume is created, with the dimensions of the sweep. Here, the length of the z-axis is set to the length of the sweep. For each pixel classified as lumen, the 3-D coordinates are calculated according to the probe tracking coordinates. Next, the nearest voxel is categorized as lumen. This is repeated for every pixel. The voxel volumes are visualized using an isosurface representation of the volume data.

III. RESULTS

A. Training

All networks were implemented in Keras [23] a Python library built upon TensorFlow [24]. The neural networks were trained using a 5-folds cross-validation approach as described in Section II-D. Figure 4 shows the average DICE of all 5 folds combined on the training and validation sets after each epoch

for each architecture. It can be observed that during training the DICE is increasing on both the training and validation set. The maximum DICE values on the validation set are obtained after 95-100 epochs. The validation loss improvements are minor so the models are considered to have converged. The differences of the maximum validation DICE between the different architectures are very small, for the training set "Carotid" 0.03% and training set "All" 0.15%, respectively. The maximum validation DICE for networks trained with carotid data (0.971) are higher compared to networks trained with all training data (0.953- 0.955). The standard deviation of the Dense U-net architecture is higher compared to the other architectures.

B. Segmentation performance on test sets

Table IV clarifies the different test sets that are used in this section. Table V shows the average Recall and Precision (of the 5 folds) on Test Set 1. Here the networks trained with training set "All" perform better in terms of Recall than the networks trained with Carotid data. However, this is at the expense of the Precision, which is higher for the networks trained with only carotid data. The Multires U-net shows the best trade-off in Recall and Precision. Table VI shows the DICE, IoU, Recall and Precision values for each network on Test Set 2. The Multires U-net trained with "All" data achieved the highest average DICE of 0.923 and IoU of 0.870. The Multires U-net shows high consistency over all folds (low standard deviation). The maximum DICE of a single fold was 0.940, achieved by the Compact U-net trained on carotid data. The DICE and IoU scores of the Dense U-net are significant lower compared to the other architectures. This is caused by a low Precision, thus there are a high number of false positive predictions, whilst the Recall is still acceptable. Figure shows 5 samples from Test Set 2 with corresponding boundaries of the segmentation performed by the different architectures. The first three images are selected to showcase segmentation errors. The first example (left) shows an artery shaped, false positive prediction by the Dense U-net and the Residual U-net is not able to detect an artery (false negative). In the second image the background shows irregular shaped objects, falsely detected by the Dense and Compact U-net. Image 3 shows over-estimation of the vessel area by the Compact and Multires U-net. A video of the segmentation results on Test Set 2 using the Multires U-net is shown in the Supplementary Materials. Table VII shows the performance on Test Set 3. The average DICE coefficients are all in the range between 0.927-0.932. Overall, the networks trained on Carotid data achieve a slightly higher DICE (0.3%) compared when trained on "All" data. The maximum DICE of a single fold was 0.9340, achieved by the Residual U-net trained on "All" data. All trained architectures show a very high Recall and Precision, in many cases there were only true positive predictions and no false negative and false positive predictions (This is shown by a Recall and Precision value of 1.000). Overall higher DICE, IoU, Recall and Precision are obtained on Test Set 3 compared to Test Set 2. Test Set 2 consists of images obtained with the same US set-up (US device + transducer) as used for the training data. Figure 6 shows the DICE distribution for architectures trained with training set "Carotid" and "All". Here the DICE was calculated for every architecture for the five folds and scores were combined in the histogram. The DICE distributions on Test Set 3 of "Carotid" and "All" are similar, the maximum difference for a bin is 0.6% of the total set. Test Set 2 shows more DICE scores between 0.00-0.30 and 0.95-1.00 for architectures trained with "Carotid". Table VIII shows the comparison between segmentation of Test Set 3, using the Compact U-net and the semi-automatic method that was used to generate the masks. The images from Test Set 3 were sampled from ten different B-mode sequences. In each B-mode sequence a manual seed was placed at the first frame of the CCA. The images from Test Set 3 were selected and the metrics were calculated. The semi-automatic algorithm was able to track the arteries in all cases, showing a recall and precision of 1.000. With a DICE of 0.922 and IOU of 0.858 the semi-automatic algorithm performs slightly worse than the different U-net architectures.

TABLE V
THE RECALL AND PRECISION WERE CALCULATED ON TEST SET 1 FOR BOTH TRAINING SETS. THE AVERAGE VALUES OVER 5-FOLDS ARE SHOWN.

Test Set 1	Training set: Carotid		Training set: All	
Architecture	Recall	Precision	Recall	Precision
Standard	0.987	**0.987**	**1.000**	0.948
Compact	0.992	0.828	0.999	0.876
Dense	0.957	0.478	0.998	0.598
Residual	0.991	**0.987**	0.998	0.942
Multires	0.997	0.985	0.999	0.960

Table IX shows the performance of the different architectures on Test Set 4. The networks trained with dataset "Carotid" show poor performance (DICE < 0.4) on this test set, which is expected since these networks are not trained with arm data. When trained with dataset "All", the Standard U-net performed best on this test set in terms of DICE, IoU and Recall. The Multires has the highest Precision (0.860). Compared to Test Set 2 & 3 the performance is worse.

C. Segmentation performance on test sets per epoch.

In this study, the test sets are different from the training and validation data, i.e., the test sets are manual annotated and different US devices were used. It could be that the best performing network on the validation data does not result in the best performance on the test sets. Figure 7 shows the DICE evaluated after each epoch on Test Set 2 and 3. On average, the maximum DICE was found after 11 epochs for Test Set 2 (x) and 33 epochs for Test Set 3(*). After this, the DICE slowly decays with increasing epochs, this in contrary to the DICE on the validation data, that kept increasing each epoch (Figure 4). This decay is more pronounced on Test Set 2 compared to Test Set 3. Although this decay is also present in the Multires U-net, this architecture shows the most consistent performance on Test Set 2 and 3. The Multires U-net gives the overall maximum DICE of 0.943 on Test Set 2, when trained on "All" data, whereas the Standard U-net reaches the overall maximum DICE of 0.940 on Test Set 3 (also trained on "All" data).

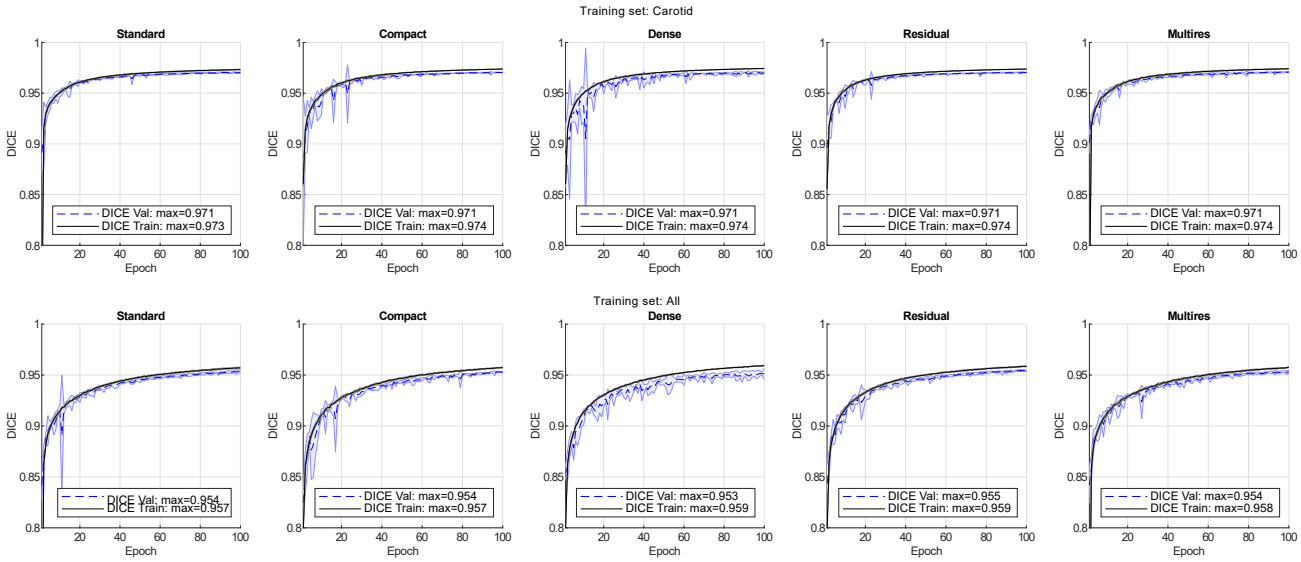

Fig. 4. The DICE per epoch during training is shown for the training set (black) and validation set (dashed blue). This is the average DICE of the 5-folds, the standard deviation is shown by the error bars. The legend shows the maximum average DICE over all epochs for the training and validation set.

TABLE VI
THE AVERAGE DICE, IOU, RECALL AND PRECISION WITH THE CORRESPONDING STANDARD DEVIATION OF THE 5 FOLDS WERE CALCULATED ON TEST SET 2. THE RESULTS ARE SHOWN FOR TWO SETS OF TRAINING DATA: "CAROTID" AND "ALL".

Test Set 2	Training set: Carotid				Training set: All			
Architecture	DICE	IoU	Recall	Precision	DICE	IoU	Recall	Precision
Standard	0.912 ± 0.014	0.863 ± 0.013	0.966 ± 0.018	0.982 ± 0.010	0.911 ± 0.013	0.859 ± 0.013	0.968 ± 0.013	0.973 ± 0.010
Compact	0.915 ± 0.019	0.863 ± 0.020	0.974 ± 0.020	0.901 ± 0.113	0.882 ± 0.058	0.821 ± 0.072	0.978 ± 0.019	0.899 ± 0.126
Dense	0.785 ± 0.051	0.707 ± 0.056	0.964 ± 0.031	0.486 ± 0.105	0.862 ± 0.074	0.796 ± 0.090	**0.990 ± 0.005**	0.680 ± 0.227
Residual	0.916 ± 0.028	0.866 ± 0.025	0.971 ± 0.031	0.974 ± 0.020	0.906 ± 0.023	0.853 ± 0.023	0.966 ± 0.024	0.966 ± 0.029
Multires	0.919 ± 0.016	0.869 ± 0.016	0.977 ± 0.015	**0.987 ± 0.004**	**0.923 ± 0.008**	**0.870 ± 0.007**	0.986 ± 0.010	0.982 ± 0.012

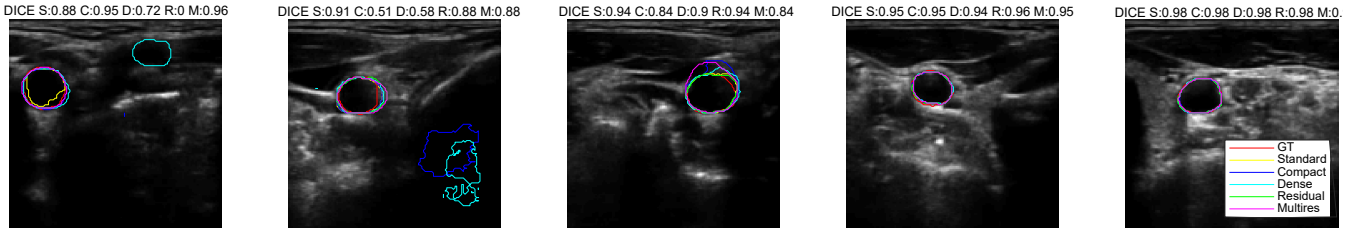

Fig. 5. Five images of Test Set 2, including the boundaries of the segmentation of the 5 architectures, fold 1. The corresponding ground truth is shown in red. The DICE for each segmentation is displayed in the title for the Architectures: Standard (S), Compact (C), Dense (D), Residual (R) and Multires (M). The networks were trained with training set "All". The first three images are selected based on different errors.

TABLE VII
THE AVERAGE DICE, IOU, RECALL AND PRECISION WITH THE CORRESPONDING STANDARD DEVIATION OF THE 5 FOLDS WERE CALCULATED ON TEST SET 3. THE RESULTS ARE SHOWN FOR TWO SETS OF TRAINING DATA: "CAROTID" AND "ALL".

Test Set 3	Training set: Carotid				Training set: All			
Architecture	DICE	IOU	Recall	Precision	DICE	IOU	Recall	Precision
Standard	0.928 ± 0.008	0.872 ± 0.009	0.994 ± 0.010	0.994 ± 0.014	0.927 ± 0.008	0.869 ± 0.007	0.996 ± 0.006	0.996 ± 0.010
Compact	**0.932 ± 0.001**	**0.875 ± 0.002**	**1.000 ± 0**	0.998 ± 0.005	0.927 ± 0.008	0.869 ± 0.008	0.985 ± 0.017	0.996 ± 0.010
Dense	0.928 ± 0.007	0.870 ± 0.009	0.996 ± 0.006	0.967 ± 0.068	0.927 ± 0.007	0.869 ± 0.008	0.987 ± 0.011	0.996 ± 0.006
Residual	**0.932 ± 0.001**	0.872 ± 0.001	**1.000 ± 0**	**1.000 ± 0**	0.930 ± 0.005	0.872 ± 0.006	0.998 ± 0.005	0.998 ± 0.005
Multires	0.930 ± 0.006	0.874 ± 0.008	0.998 ± 0.005	0.992 ± 0.014	0.927 ± 0.007	0.869 ± 0.006	**1.000 ± 0**	0.996 ± 0.010

TABLE VIII
THE PERFORMANCE OF THE SEMI-AUTOMATIC METHOD ON TEST SET 3 VERSUS THE COMPACT U-NET TRAINED WITH TRAINING SET "CAROTID".

Test Set 3	DICE	IOU	Recall	Precision
Semi-automatic method[12]	0.922	0.858	1.000	**1.000**
Compact U-net	**0.932**	**0.875**	**1.000**	0.998

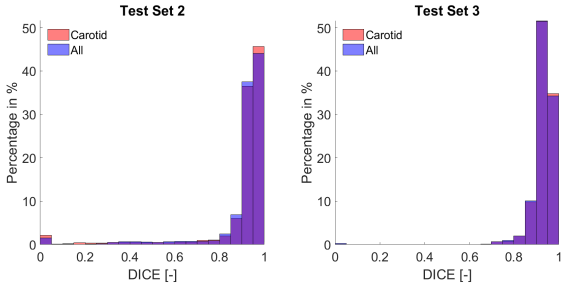

Fig. 6. The DICE histogram for architectures trained with the "Carotid" and "All" datasets on Test Set 2 and Test Set 3. All DICE results of the 5 architectures and 5 folds were combined in these histograms.

D. Reconstruction Geometries in 3-D

Figure 8 shows a 3-D geometry of the carotid artery including the bifurcation. The segmentation was performed using the Compact architecture, trained with "All" training data. Acquisitions of the arm of two volunteers were left out of the training set. Three sweeps were made along the cephalic vein, radial artery, and the ulnar artery. Figures 9 and III-D show the geometries reconstructed from each sweep, and the combined B-mode data. The images were segmented using the Compact architecture. Each sweep consisted of approximately 900 images.

IV. DISCUSSION

In this study we have presented a generalized method to segment the lumen-wall boundary in transverse US images of central and peripheral blood vessels using CNNs. Semi-3D data were acquired using freehand sweeps while performing 2-D US imaging. The training data were labeled by a semi-automatic segmentation algorithm [12] and the lumen-wall boundary segmentation was visually checked before inclusion to the training set. The semi-automatic segmentation method was developed for sequences of transverse US images of the carotid artery and requires manual input, whereas the method proposed is fully automatic and works on stand-alone images (or sequences) of different arteries and veins of different sizes and at different depths. To ensure the semi-automatic method's performance on images of the forearm and leg, multiple starting seeds were selected manually, thereby avoiding tracking errors. Training CNNs requires large datasets, however, in the field biomedical engineering training data are not often widely available. In order to create a larger database, data of different arteries and veins were combined. The jugular vein next to the carotid artery was not included, since it is most of the times compressed or/and has a non-circular shape, and also small blood vessels were neglected. Small blood vessels in the image could result in false positive predictions, however these predictions could be easily removed in post-processing since they lack continuity in the sweeps. Five state-of-the-art U-net architectures were tested and trained in 5-folds cross-validation with two different training sets: transverse US images of the carotid artery, and transverse US images of multiple arteries. We found that the addition of images of different blood vessels (radial artery, ulnar artery, femoral artery and cephalic vein) to the training set not only enabled segmentation of different types of arteries, but also slightly improved the general performance of the neural network for the segmentation tasks of the carotid artery, despite differences in size, structure, and surrounding tissues of the input training sets provided. The different architectures all converged when training on the training/validation sets and did not show signs of overfitting (both training loss and validation loss kept decreasing). The performance of the segmentation is tested using an online CCA database (Test Set 1 and 2) and two in-house datasets (Test Set 3 and 4), an overview is given in Table IV. Modification of the building blocks of the U-net shown in 3 have a small influence on the performance of the network compared to the Standard U-net on the test sets. However, the Multires architecture shows the highest training consistency, revealing a high performance of all folds on all test sets. In this study, the Dense U-net showed precarious results on Test Set 1 and 2, showing a very low DICE, IoU and Precision. Whilst the performance on Test Set 3 and 4 were comparable to the other architectures. Thus this network performed insufficient on images made by a different US machine than those in the training set. However, it should be noted that the hyper-parameters were not optimized for each network separately.

It is observed (Tables V, VI) that the inclusion of all blood vessels to the training set improves the detection of the CCA (higher Recall), however, this comes with a lower Precision. This means that there are more false positive predictions, which could be caused by the nature of the arm and leg images in the "All" vessel training set, since they contain multiple arteries/ veins in one image. These false positive predictions cause the slightly lower DICE and IOU scores for networks trained with "All". This can also be seen in Figure 6, where there are less low DICE scores between 0.00-0.30, which indicate better detection. But there are also lower scores between 0.95-1.00 when trained with "All". In terms of the DICE and IoU we do not see differences between the two training sets. However, when Test Set 2 & 3 are evaluated after each epoch (Figure 7), the highest DICE on Test Set 2 and 3 are achieved by networks trained with "All" vessel data (DICE of 0.943 and 0.940, respectively). This supports the conclusion that could be beneficial to add data of other arteries for segmentation purposes, especially when limited training data are available.

The training/validation sets could not be used as test set, since the masks were semi-automatically generated, and even though the bad samples were removed, this cannot be considered as ground truth. Figure 7 shows the evaluation of the DICE after each epoch on Test Set 2 and 3. For all configurations the maximum DICE was already found after 10-40 epochs, whilst for the validation sets this was always between 90-100 epochs (Figure 4). This indicates that the

TABLE IX
THE AVERAGE DICE, IOU, RECALL AND PRECISION WITH THE CORRESPONDING STANDARD DEVIATION OF THE 5 FOLDS WERE CALCULATED ON TEST SET 4.

Test Set 4	Training set: Carotid				Training set: All			
Architecture	DICE	IOU	Recall	Precision	DICE	IOU	Recall	Precision
Standard	0.381 ± 0.040	0.352 ± 0.037	0.412 ± 0.043	0.833 ± 0.076	**0.759 ± 0.009**	**0.671 ± 0.014**	**0.874 ± 0.017**	0.837 ± 0.036
Compact	0.396 ± 0.033	0.363 ± 0.030	0.424 ± 0.044	0.775 ± 0.047	0.735 ± 0.021	0.661 ± 0.015	0.840 ± 0.040	0.845 ± 0.030
Dense	0.346 ± 0.067	0.313 ± 0.062	0.384 ± 0.073	0.600 ± 0.254	0.731 ± 0.017	0.662 ± 0.011	0.842 ± 0.016	0.825 ± 0.035
Residual	0.335 ± 0.027	0.303 ± 0.025	0.360 ± 0.032	0.771 ± 0.057	0.741 ± 0.019	**0.671 ± 0.019**	0.848 ± 0.018	0.830 ± 0.044
Multires	0.374 ± 0.047	0.343 ± 0.045	0.408 ± 0.053	0.783 ± 0.072	0.745 ± 0.015	0.657 ± 0.019	0.848 ± 0.016	**0.860 ± 0.011**

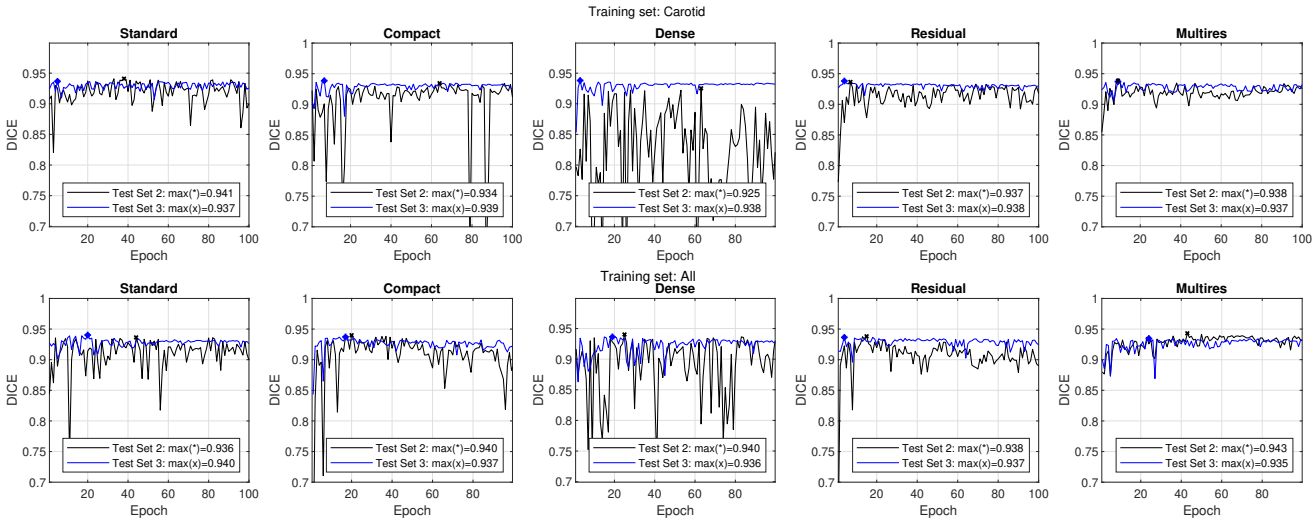

Fig. 7. The DICE evaluated after each epoch on Test Set 2 (black) and Test Set 3 (blue). The maximum average DICE is marked with (*) for Test Set 2 and a (x) for Test Set 3.

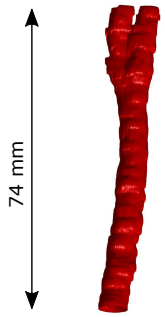

Fig. 8. Example: Automatically created 3-D geometry of the carotid artery including CCA, ICA and ECA.

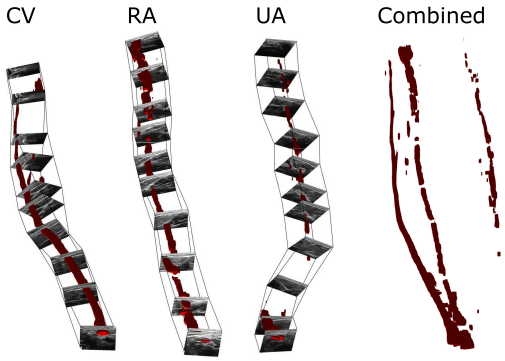

Fig. 9. Example 1: lower arm vasculature, including the cephalic vein, radial artery, and the ulnar artery. Each sweep is displayed separately with a subset of the US images overlaid.

model is learning complex features after 10-40 epochs that are specific for the training/validation sets, but are not general features (i.e. specific for US machine (Test Set 2) or for the semi-automatic method (Test Set 3)). Thus when aiming for a more generalized segmentation model of blood vessels, it could be beneficial to not fully converge the training, even though this would yield better results on the validation set.

The performance is compared with literature using an online CCA database (Test set 1 and 2). On Test Set 1, Recall of 1.000 was achieved by multiple architectures trained using "All" vessels. Literature shows an accuracy of 88% and a detection rate of 90% on 433 images from the same database [5]. A detection rate of 97% was reported by [17]. A DICE of 0.94 and an IoU of 0.88 were achieved on Test Set 2 and

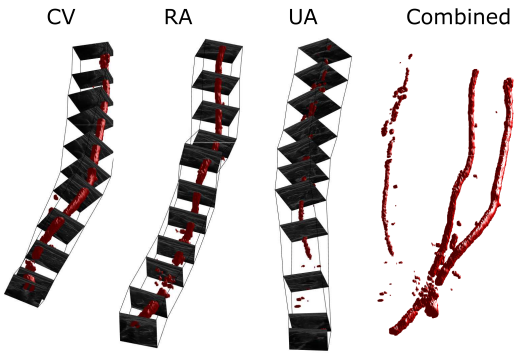

Fig. 10. Example 2: lower arm vasculature, including the cephalic vein, radial artery, and the ulnar artery. Each sweep is displayed separately with a subset of the US images overlaid.

3. This is higher when compared to literature: several state-of-the-art studies report a DICE of 0.91 on healthy CCA US images, however, the validation analyses were not performed on the same datasets [25]. The semi-automatic method that was used to generate the masks showed a DICE of 0.91 on the CCA in the original paper [12]. On Test Set 3 a DICE of 0.922 was found, which is slightly lower than the different U-net architectures on Test Set 3.

Large area 3-D geometries of arteries and veins are constructed, based on 2-D transverse B-mode sequences in combination with a probe tracking device. Figure 8 shows the geometry of a carotid artery including the bifurcation. The pulsation of arterial wall is visible., which could be removed in post processing. Figures 9 and III-D show the obtained 3-D geometries of the RA, UA, and CV in the lower arm of two volunteers. Upon visual inspection, the cephalic vein and radial artery segmentation performances are good. The ulnar artery is smaller and lies deeper in the image, so the image quality is less optimal, or the ulnar artery is not visible in the image at all. This causes the gaps in the 3-D geometry. The segmentation results on Test Set 4 were less accurate than on the other test sets. An average DICE of 0.759 was found using the Standard U-net. It should be noted that annotations only contain segmentations of the artery/vein that was followed during the acquisition, since it was not possible to correctly manually segment all vessels present in the image. Furthermore, the acquisition itself could be improved in future studies by developing more dedicated measurement protocols for the different applications.

One study reported 3-D geometry assessment of radio-cephalic arteriovenous fistula, a connection between the cephalic vein and radial artery [26], [27]. Here, the segmentation was performed with commercial software and required manual interventions. Prevost, et al. [28] showed 3-D reconstruction of sweep data without external probe tracking, but by using a CNN to estimate the motion of successive ultrasound frames. This creates new future opportunities, also for large clinical trials, using existing equipment and without the use of a probe tracking device. Combining CNN-based probe tracking with our segmentation could generate 3-D geometries directly from B-mode sequences without any dedicated, hardware-based probe tracking. The geometries presented in this study, display direct results of the segmentation, without further post-processing or cosmetics. To create a final 3-D anatomical map, several steps could be taken such as centerline detection, heartbeat removal and spatial filtering [29], [26], [12]. However, the validity and applicability of such post-processing steps depend on the final application.

Unfortunately, there was no ground truth data available for 3-D geometry validation. This would require contrast angiography imaging using CT or MRI, which would be an extra burden to the volunteers, and would have to be executed in the clinic. In a future study, these verification measurements could be possible using clinically available multi-modality datasets of patients. This is beyond the scope of the present study, which is aimed at investigating the feasibility and merit of training a CNN on a generalized dataset for segmentation of stacks of 2-D vascular US images.

In this study, single frame-by-frame segmentations were performed, yet the possible benefit of using multiple frames, or a 3-D reconstructed volume frame as input remains unexplored. A fully annotated dataset (manual segmentation) would be preferable for the training of the network. However, this study shows that the segmentation results on manual segmented test sets are good, despite the use of semi-automatic annotated masks for the training data. The segmentation method and results of previous studies were used as training dataset, since these segmentations were already validated, and to reduce manual labor. In addition, all frame-by-frame results were checked by trained observers to improve the accuracy of the training set further.

V. Conclusion

A CNN was successfully deployed for the segmentation of stacks of transverse Ultrasound B-mode images of different arteries in vivo. Five different configurations of the U-net architecture were tested. Overall, the Multires U-net architecture showed the most consistent performance, with DICE = 0.93 and IoU = 0.88 on 269 manual segmented images of an online CCA database. A Recall of 0.999 and a Precision of 0.960 were found on 971 images. The addition of images of different arteries can be a valuable strategy to generate more data to train a CNN for the task of segmenting a single artery despite different size, structure, surrounding tissue etc. It enables generalized segmentation of different vessels with a single network, whilst maintaining the segmentation quality. With this set-up, large area vasculature can be imaged in the extremities, automatically segmented, and converted into a 3-D geometry for diagnosis or intervention planning. Future work will focus on patient studies and validation of the method proposed.

ACKNOWLEDGMENT

This study was funded by Stichting Lijf and Leven. This work was carried out on the Dutch national e-infrastructure with support by SURF Cooperative.

REFERENCES

[1] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[2] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[3] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[4] E. Smistad and L. Løvstakken, "Vessel detection in ultrasound images using deep convolutional neural networks," in *Deep Learning and Data Labeling for Medical Applications*. Springer, 2016, pp. 30–38.

[5] P. K. Jain, S. Gupta, A. Bhavsar, A. Nigam, and N. Sharma, "Localization of common carotid artery transverse section in b-mode ultrasound images using faster rcnn: a deep learning approach," *Medical & Biological Engineering & Computing*, pp. 1–12, 2020.

[6] S. S., J. K. B., R. C., N. Madian, and S. T., "Convolutional neural network for segmentation and measurement of intima media thickness," *Journal of Medical Systems*, vol. 42, no. 8, p. 154, Jul 2018. [Online]. Available: https://doi.org/10.1007/s10916-018-1001-y

[7] K. Lekadir, A. Galimzianova, À. Betriu, M. del Mar Vila, L. Igual, D. L. Rubin, E. Fernández, P. Radeva, and S. Napel, "A convolutional neural network for automatic characterization of plaque composition in carotid ultrasound," *IEEE journal of biomedical and health informatics*, vol. 21, no. 1, pp. 48–55, 2016.

[8] R. Zhou, A. Fenster, Y. Xia, J. D. Spence, and M. Ding, "Deep learning-based carotid media-adventitia and lumen-intima boundary segmentation from three-dimensional ultrasound images," *Medical physics*, vol. 46, no. 7, pp. 3180–3193, 2019.

[9] M. Xie, Y. Li, Y. Xue, R. Shafritz, S. A. Rahimi, J. W. Ady, and U. W. Roshan, "Vessel lumen segmentation in internal carotid artery ultrasounds with deep convolutional neural networks," in *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2019, pp. 2393–2398.

[10] E. Smistad, K. F. Johansen, D. H. Iversen, and I. Reinertsen, "Highlighting nerves and blood vessels for ultrasound-guided axillary nerve block procedures using neural networks," *Journal of Medical Imaging*, vol. 5, no. 4, p. 044004, 2018.

[11] D. Mishra, S. Chaudhury, M. Sarkar, and A. S. Soin, "Ultrasound image segmentation: a deeply supervised network with attention to boundaries," *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 6, pp. 1637–1648, 2018.

[12] J. de Ruijter, M. van Sambeek, F. van de Vosse, and R. Lopata, "Automated 3d geometry segmentation of the healthy and diseased carotid artery in free-hand, probe tracked ultrasound images," *Medical Physics*, vol. 47, no. 3, pp. 1034–1047, 2020.

[13] B. Chiu, A. Krasinski, J. D. Spence, G. Parraga, and A. Fenster, "Three-dimensional carotid ultrasound segmentation variability dependence on signal difference and boundary orientation," *Ultrasound in medicine & biology*, vol. 36, no. 1, pp. 95–110, 2010.

[14] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual u-net," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 5, pp. 749–753, 2018.

[15] M. Kolařík, R. Burget, V. Uher, K. Říha, and M. K. Dutta, "Optimized high resolution 3d dense-u-net network for brain and spine segmentation," *Applied Sciences*, vol. 9, no. 3, p. 404, 2019.

[16] N. Ibtehaz and M. S. Rahman, "Multiresunet: Rethinking the u-net architecture for multimodal biomedical image segmentation," *Neural Networks*, vol. 121, pp. 74–87, 2020.

[17] K. Říha, J. Mašek, R. Burget, R. Beneš, and E. Závodná, "Novel method for localization of common carotid artery transverse section in ultrasound images using modified viola-jones detector," *Ultrasound in medicine & biology*, vol. 39, no. 10, pp. 1887–1902, 2013.

[18] P. Abolmaesumi, S. E. Salcudean, W.-H. Zhu, M. R. Sirouspour, and S. P. DiMaio, "Image-guided control of a robot for medical ultrasound," *IEEE Transactions on Robotics and Automation*, vol. 18, no. 1, pp. 11–23, 2002.

[19] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *International journal of computer vision*, vol. 1, no. 4, pp. 321–331, 1988.

[20] D. Müller and F. Kramer, "Miscnn: A framework for medical image segmentation with convolutional neural networks and deep learning," 2019.

[21] F. Isensee, P. Jäger, J. Wasserthal, D. Zimmerer, J. Petersen, S. Kohl, J. Schock, A. Klein, T. Roß, S. Wirkert, P. Neher, S. Dinkelacker, G. Köhler, and K. Maier-Hein, "batchgenerators - a python framework for data augmentation." 2020.

[22] H. Chen, X. Qi, L. Yu, Q. Dou, J. Qin, and P.-A. Heng, "Dcan: Deep contour-aware networks for object instance segmentation from histology images," *Medical image analysis*, vol. 36, pp. 135–146, 2017.

[23] F. Chollet *et al.*, "Keras," https://keras.io, 2015.

[24] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, software available from tensorflow.org. [Online]. Available: https://www.tensorflow.org/

[25] E. Smistad and F. Lindseth, "Real-time automatic artery segmentation, reconstruction and registration for ultrasound-guided regional anaesthesia of the femoral nerve," *IEEE Transactions on Medical Imaging*, vol. 35, no. 3, pp. 752–761, 2015.

[26] J. E. Carroll, E. S. Colley, S. D. Thomas, R. L. Varcoe, A. Simmons, and T. J. Barber, "Tracking geometric and hemodynamic alterations of an arteriovenous fistula through patient-specific modelling," *Computer methods and programs in biomedicine*, vol. 186, p. 105203, 2020.

[27] E. Colley, J. Carroll, S. Thomas, R. L. Varcoe, A. Simmons, and T. Barber, "A methodology for non-invasive 3-d surveillance of arteriovenous fistulae using freehand ultrasound," *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 8, pp. 1885–1891, 2017.

[28] R. Prevost, M. Salehi, S. Jagoda, N. Kumar, J. Sprung, A. Ladikos, R. Bauer, O. Zettinig, and W. Wein, "3d freehand ultrasound without external tracking using deep learning," *Medical image analysis*, vol. 48, pp. 187–202, 2018.

[29] E. Yeom, K.-H. Nam, C. Jin, D.-G. Paeng, and S.-J. Lee, "3d reconstruction of a carotid bifurcation from 2d transversal ultrasound images," *Ultrasonics*, vol. 54, no. 8, pp. 2184–2192, 2014.