Title: Deep Learning-based femoral cartilage automatic segmentation in ultrasound imaging for guidance in robotic knee arthroscopy

Authors: M. Antico[1, 2], F. Sasazawa[3], M. Dunnhofer[4], S.M. Camps[5, 6], A.T. Jaiprakash[2, 7], A.K. Pandey[2, 7], R. Crawford[1, 2], G. Carneiro*[8], D. Fontanarosa*[2, 9]

[1]School of Chemistry, Physics and Mechanical Engineering, Science and Engineering Faculty, Queensland University of Technology, Brisbane, QLD 4000, Australia

[2]Institute of Health & Biomedical Innovation, Queensland University of Technology, Brisbane, QLD

[3]Department of Orthopaedic Surgery, Faculty of Medicine and Graduate School of Medicine, Hokkaido University, Sapporo**,** Japan

[4]Department of Mathematics, Computer Science and Physics, University of Udine, Udine, Italy

[5]Faculty of Electrical Engineering, Eindhoven University of Technology, Eindhoven, the Netherlands

[6]Oncology Solutions Department, Philips Research, Eindhoven, the Netherlands

[7]School of Electrical Engineering, Computer Science, Science and Engineering Faculty, Queensland University of Technology, Brisbane, QLD 4000, Australia

[8]Australian Institute for Machine Learning, School of Computer Science, the University of Adelaide, Adelaide, Australia

[9]School of Clinical Sciences, Queensland University of Technology, Gardens Point Campus, 2 George St, Brisbane, QLD 4000, Australia

*Both authors contributed equally to this manuscript

Corresponding author: Davide Fontanarosa

Gardens Point Campus, 2 George St, Brisbane, QLD 4000, Australia

cell: +61 (0) 403862724

email: d3.fontanarosa@qut.edu.au

Abstract

Knee arthroscopy is a minimally invasive surgery used in the treatment of intra-articular knee pathology which may cause unintended damage to femoral cartilage. An ultrasound(US)-guided autonomous robotic platform for knee arthroscopy can be envisioned to minimise these risks and possibly to improve surgical outcomes. The first necessary tool for reliable guidance during robotic surgeries was an automatic segmentation algorithm to outline the regions at risk.

In this work, we studied the feasibility of using a state-of-the-art deep neural network (UNet) to automatically segment femoral cartilage imaged with dynamic volumetric US (at the refresh rate of 1 Hz), under simulated surgical conditions. Six volunteers were scanned which resulted in the extraction of 18278 2D US images from 35 dynamic 3D US scans, and these were manually labelled. The UNet was evaluated using a 5-fold cross-validation with an average of 15531 training and 3124 testing labelled images per fold. An intra-observer study was performed to assess intra-observer variability due to inherent US physical properties. To account for this variability, a novel metric concept named Dice coefficient with boundary uncertainty (DSC_{UB}) was proposed and used to test the algorithm. The algorithm performed comparably to an experienced orthopaedic surgeon, with DSC_{UB} of 0.87. The proposed UNet has the potential to localise femoral cartilage in robotic knee arthroscopy with clinical accuracy.

Keywords: Ultrasound-guided minimally invasive surgery, Ultrasound-Guided Arthroscopy, Robotic knee arthroscopy, Femoral cartilage automatic segmentation, Deep learning, Robotic knee arthroscopy navigation.

Introduction

Knee arthroscopy is a well-established minimally invasive surgery (MIS). It utilises an endoscope (for these applications referred to as "arthroscope") and surgical tools to inspect and repair knee structures (McKeon et al. 2009). During the intervention, typically the surgeon visually relies on the 2D field of view (FOV) of the surgical site (and on its depth interpretation), provided by the arthroscope and projected on a screen. While looking at the screen, the surgeon moves the leg to different flexion angles to make sufficient space for the arthroscope/tools to reach the different knee structures. The procedure can be complicated and may lead to unintended femoral cartilage damage and post-surgical complications (Jaiprakash et al. 2017; Price et al. 2015).

Surgical outcomes may improve with the use of a robotic assistive platform which provides surgical guidance by combining the arthroscopic view with real-time volumetric ultrasound (US) (Wu et al. 2019). Presently, US is the only clinical imaging modality which can intra-operatively map all the internal regions of interest in the knee joint to enable safe surgical tool guidance (Antico et al. 2019). For real-time applications such as this, it is necessary to provide automation of US image interpretation, and this has proved challenging.

In addition to known limitations such as uncontrolled speckle noise, intensity inhomogeneity and sensitivity to probe position, new sources of image variability include soft tissue deformation and the relative motion between the knee bony structures (Alves et al. 2016; Faisal et al. 2018c; Shrimali et al. 2009; Vlad 2015). For these reasons, clinical applications of US imaging for the knee joint are presently limited to pathology diagnosis and percutaneous needle injections, and these are highly operator dependent (Alves et al. 2016; Cianca et al. 2014; Grzelak et al. 2016; Lueders et al. 2016; Paczesny and Kruczyński 2011; Razek et al. 2009).

The first step to enable US guidance for robotic knee arthroscopy is a system that can automatically outline the regions at risk in the US images. Recently, deep learning (DL) algorithms and in particular convolutional neural networks (CNN) have been introduced for image segmentation in medical imaging. CNN outperforms traditional imaging processing approaches, especially in cases

where a high degree of variability in the image quality and appearance is present (Huang et al. 2018). These types of algorithms have been employed for US images of different body regions, for example, segment vessels (Smistad and Løvstakken 2016), fetal abdomen (Ravishankar et al. 2016), brain regions (Milletari et al. 2017) and tongue (Jaumard-Hakoun et al. 2016). DL algorithms have not yet been utilised for US images of the knee, to the best of our knowledge.

The femoral cartilage is most commonly damaged during knee arthroscopy (Curl et al. 1997; Jaiprakash et al. 2017). Therefore, in this work, we study the feasibility of using a state-of-the-art CNN (UNet) (Ronneberger et al. 2015) to segment this structure automatically. The anterior aspect of the knee of the six volunteers was imaged using volumetric dynamic US, or 4D US (3D+time). This dynamic image acquisition was performed under simulated surgical conditions, with the leg flexed or the probe slightly translated from the original position.

While several real-time, traditional (non-deep learning) image segmentation algorithms have been developed for US images of knee bony structures and vessels (Aka et al. 2017; Guerrero et al. 2007; Kowal et al. 2007), there seems to be very limited interest in the segmentation of the femoral cartilage within the current literature. This is due to the limited use of US for knee cartilage scans, currently performed only for diagnostic purposes. Faisal et al. (2018a) compared different types of level sets algorithms to segment the cartilage layer in US images for automatic cartilage thickness computation in diagnostic procedures. In that study, 2D US images were collected and used to test the algorithm. The US images were acquired considering only one leg position, with the knee in full extension and the probe placed along the patient's medial-lateral direction below the patella. With this type of scanning technique, the cartilage is imaged only along the plane where the sound waves are perpendicular to the cartilage layer, making its boundaries appear highly hyperechoic (very bright) (*Figure 1*).

In contrast, our imaging is fully volumetric, so this highly hyperechoic condition occurs only for a part of the imaged femoral cartilage, and this results in the cartilage boundaries not often being clearly visible. For our application, it is paramount for the algorithm to be able to outline the cartilage

layer in addition to those areas where the exact location of its boundaries is uncertain. This introduces additional challenges not only in terms of image segmentation but also for the evaluation of the algorithm performance. In such cases, standard metrics do not correctly evaluate the algorithm's performance as they might penalise parts of the segmentation that may belong to the actual target. To mitigate for this, a revised version of the Dice similarity coefficient was developed which accounts for the uncertainty in the ground-truth boundaries. An experienced orthopaedic surgeon re-contoured a subset of the whole dataset to assess the uncertainties as an intra-observer study. The performance of the UNet and expert was then evaluated and compared using a 5-fold cross-validation with an average of 15531 training and 3124 testing labelled images per fold, considering both the standard DSC and the introduced novel metric.

Materials and Methods

Image data acquisition

Data collection was performed by F.S. using a state-of-the-art clinical system (Philips EpiQ7 US system, Philips Medical Systems, Andover, MA, United States) and a 2D US probe (Philips VL13-5 US transducer, Philips Medical Systems, Andover, MA, United States). The Queensland University of Technology Ethics Committee granted the approval for data acquisition (No. 1700001110). An US specialist optimised the workstation settings for knee structure visualization: 13 MHz probe frequency; 4 cm penetration depth; far field focus; dynamic range of 60 dB; emission power of -0.5 dB and medium persistence. Whenever needed, penetration depth, focus and dynamic range were adjusted based on the volunteer variations/characteristics and the angle of knee flexion during the scanning procedure. The penetration depth varied between 3.5 and 6 cm and the dynamic range between 48 and 60 dB. SonoCT real-time compound imaging technology and XRES image processing were selected on the US system during the image acquisition to enhance image quality.

Considering the constraints imposed by the presence of surgical tools through the medial/lateral patella portals, the images were acquired with the probe placed on the patella tendon, parallel to the principal axis of the tibia. With this probe positioning, the edges of the transducer surface may not have complete contact with the skin due to the irregularities on the surface of the knee. These gaps at the interface between the knee surface and the probe delimited the knee region visible in the US volumes and caused an overall image quality degradation. To capture images with the maximum possible FOV, and to minimise these acoustic coupling artefacts, scanning was completed with the volunteers submerged in water (*Figure 2*).

Three types of scans were performed covering all the possible surgical scenarios with a leg flexion between 0 and 30 degrees. The 0 degree flexion angle was set at the "neutral" leg position (or extended leg), where the tibial tubercle and the femoral shaft were aligned (Zarins et al. 1983). The 30 degrees flexion angle was achieved by bending the leg such that the angle between the femur and the tibia with respect to the neutral leg position was 15 degrees. Fast, precise and consistent image acquisition was attained with the use of a leg cushion designed to support the leg at this flexion angle.

The US datasets were collected:

- during leg extension from 30 to 0 degrees (namely *Extension 30*);
- keeping the leg static at either 0 or 30 degrees while moving the US probe in the inferior direction from the patella tip up to the point where the femoral condyles were no longer visible (namely *Translation 0/ Translation 30*).

The scanning convention established during knee extension was imaging the area from the inferior end of the patella to the superior end of the tibia in the sagittal plane and both sides of femoral condyles in the transverse plane. Volumes of approximately (4 x 4 x 3) cm^3 were obtained with a 1 Hz full volume refresh rate.

Data description

Informed consent was obtained from all volunteers before imaging. A total of 35 4D US sequences, comprising 151 3D US volumes of the anterior knee aspect, were collected from six

volunteers. Five volunteers had healthy femoral cartilages and femoral cartilage pathology (i.e. femoral cartilage partial-thickness degeneration) was identified on the US images of one volunteer (*Table 1*). The cartilage defect was detected in five out of six 4D US sequences, and on ten 2D US slices per US volume on average.

Label generation

Orthopaedic surgeon F.S. outlined the cartilage contours on all the US volumes acquired using a graphical user interface (GUI) created specifically for this purpose in Mevislab (MeVis Medical Solutions AG, Germany) (*Figure 3*). Using the GUI, the surgeon was able to scroll through each US volume along the volunteers' sagittal, axial and coronal directions. The contours were drawn on the sagittal slices of the US volumes (highest resolution plane), with simultaneous visualisation on the other two projections. As an additional tool, the visualisation of the projection of the contours drawn on the neighbouring slices could be enabled (*Figure 3*).

Pre-processing of images and labels

To create the dataset for the UNet, the US images and the corresponding contours were pre-processed using MATLAB (Version 9.3.0 (R2017b), The Mathworks Inc. Natick, MA, United States). The 3D US volumes were sliced to 2D images along the sagittal axis, and the slices where the cartilage was not outlined were discarded from the dataset. The selected 2D images (and corresponding contours) were then rescaled by converting the image pixel dimensions to the largest ones within the dataset (pixel dimension along the image height 0.0922 mm and along the width 0.141 mm). Padding with black pixels was then applied to the images to match the size of the largest image in the set (510 pixels x 272 pixels). Finally, the images were down-sampled (304 pixels x 160 pixels) while preserving the image aspect ratio to enable faster computation.

UNet dataset creation

After pre-processing, a dataset of 18278 2D US labelled images was collected. The dataset (hereinafter referred to as *Dataset 1*) was divided into training, validation and test sets using a 60:20:20% split of the total number of volunteers, resulting in four volunteers allocated for training and one volunteer each for validation and testing. Since the 2D labelled images partition was not uniform across the different volunteers (*Table 2*), an optimisation procedure based on simulated annealing (Kirkpatrick et al. 1983) was implemented to ensure that the training, validation and test sets would contain approximately the same percentage of images per type of scan. Following the same optimisation method, four additional data distributions (*Datasets 2-5*) were generated (each time with a different volunteer in the test set) to allow for the cross-validation of the UNet hyperparameters (*Table 2*).

UNet architecture and implementation

The UNet is a state-of-the-art CNN for automatic semantic segmentation of medical images (Ronneberger et al. 2015). The model is composed of an encoder that extracts coarse features of the image input and a decoder that projects the features learned to the pixel space, recovering the original image resolution (*Figure 4*).

The encoder utilised consisted of five blocks, each composed of two sequential (3x3) convolutional layers followed by batch normalisation (Ioffe and Szegedy 2015), ReLu activation and dropout fraction (Srivastava et al. 2004) of 0.1. The encoder blocks were followed by (2x2) max pooling, halving each time the width and height of the activations. The decoder was built in a similar fashion, by replacing the max pooling layers with transposed convolutions. As in the original UNet implementation, the convolutional layers in the decoder part were concatenated with a copy of the image features output by the equivalent encoder blocks (*Figure 4*). This was performed to preserve the features learned in the production of fine-grained segmentations, which otherwise would be lost as the features pass through max pooling layers. The model architecture was implemented in Python using the PyTorch library (Paszke et al. 2017).

The model was trained for 31 epochs with mini-batches of eight images, using the Adam optimiser (Kingma and Ba 2015) with a learning rate of 10^{-3} and momentum of 0.95. As in Milletari et. al., 2016 (Milletari et al. 2016), we selected the Dice loss as cost function, with a weight decay of 10^{-5} as a regulatory term. The hyperparameters noted above were estimated based on the model performance in the validation set of the original data distribution (*Dataset 1*) and kept fixed during training of the *Datasets 2-5*. The model performance was assessed using the test sets of each data distribution *Datasets 1-5*, named *Test sets 1-5* respectively (*Table 2*). The training and testing procedures were performed on a Linux cluster with an NVIDIA Tesla P100 GPU (NVIDIA, Santa Clara, CA, USA).

Intra-observer variability tests

The expert variability was studied by performing a blinded experiment in which the surgeon re-contoured one volume in each test set in *Table 2* (*Test sets 1-5*) named *Volumes 1-5* respectively (*Table 3*). The volumes were randomly selected, ensuring that the type of scan and volume frame number within the 4D sequence (e.g. volume frame n is in the n^{th} volume recorded in chronological order within the 4D sequence) would vary among different volunteers.

UNet and intra-observer performance evaluation

In sections *Evaluation 1*, *Evaluation 2* and *Evaluation 3*, three different evaluation methods for the UNet and the intra-observer performance are reported. In these evaluations a standard and a novel metric introduced in *Evaluation metrics* were used.

Evaluation metrics

The DSC (Dice 2006) is a standard metric to measure the overlap between two binary masks: in our case, given a 2D US image *j*, the ground-truth segmentation $M^j{}_{GT}$ and the mask predicted $M^j{}_P$ by the UNet.

As described in *Eq. 1*, the DSC is the ratio of the number of pixels intersecting both masks multiplied by a factor of two, divided by the sum of the number of pixels contained in each mask, such that when a complete overlap between the two masks is present, the coefficient is 1.

$$DSC = \frac{2\,|(M^{j}{}_{GT} \cdot M^{j}{}_{P})|}{|M^{j}{}_{GT}| + |M^{j}{}_{P}|} \quad (1)$$

where $\cdot$ represents the dot product (or element-wise product), $M^{j}{}_{GT}$ and $M^{j}{}_{P} \in \{0,1\}^{r \times c}$, r and c are the number of pixels in the rows and columns of the masks, $|M^{j}{}_{GT}|$ and $|M^{j}{}_{P}|$ are the number of positive elements in each of the binary matrices.

The new metric introduced in this paper is a revised version of the DSC for those cases where it is not possible to determine the exact boundary of the target to be contoured due to uncertainty present in the annotation. In these types of situations, the expert can identify an area surrounding the target that will be defined as the uncertainty margin (UM_{GT}) where the real tissue boundary is present. Therefore, multiple masks could exist with boundaries within the uncertainty margin that would be considered acceptable (*Figure 5a*.). To generate the ground-truth a mask with an exact boundary is outlined that will be defined as standard ground-truth (M_{GT}). This solution can possibility penalise the mask predicted by the UNet, as the predicted pixels lying in the uncertainty margin will be evaluated as incorrect if they are not part of the M_{GT} .

To solve this issue, the standard ground-truth was delimited with a defined uncertainty margin, where it was assumed that each pixel within that margin had the same probability of belonging to the actual cartilage boundary. If this assumption is valid, the pixels of the predicted mask within the uncertainty margin can be considered as correct predictions and therefore, need to be part of the ground-truth in the *DSC* evaluation.

Consequently, the ground-truth with boundary uncertainty $M^{j}_{GT_{BU}}$ was defined as a particular solution among the possible acceptable ground-truth masks for a given 2D US image *j* (*Figure 5 b*.). $M^{j}_{GT_{BU}}$ can be expressed as the sum of the standard ground-truth region enclosed by the internal

boundary of the uncertainty margin $\left(I^{j}{}_{GT}\right)$, and the intersection between the uncertainty margin mask ($UM^{j}{}_{GT}$) and the prediction ($M^{j}{}_{P}$)(*Eq. 2*):

$$M^{j}_{GT_{BU}} = I^{j}{}_{GT} + UM^{j}{}_{GT} \cdot M^{j}{}_{P} \quad (2)$$

with $I^{j}{}_{GT}$, $UM^{j}{}_{GT}$and $M^{j}{}_{P} \in \{0,1\}^{r \times c}$, and *r* and *c* are the number of pixels in the rows and columns of the masks.

The Dice similarity coefficient with boundary uncertainty (DSC_{BU}) can be then formulated as the standard DSC in *Eq. 1*, replacing the standard ground-truth with the ground-truth mask with boundary uncertainty $\left(M_{GT_{BU}}\right)$ computed as in *Eq. 2*:

$$DSC_{BU} = \frac{2\,|(M^{j}_{GT_{BU}} \cdot M^{j}{}_{P})|}{\left|M^{j}_{GT_{BU}}\right| + |M^{j}{}_{P}|} \quad (3)$$

where · represents the dot product (or element-wise product) and $\left|M^{j}_{GT_{BU}}\right|$ and $\left|M^{j}{}_{P}\right|$ are the number of positive elements in each of the binary matrices.

Evaluation 1

The performance of the UNet and the intra-observer variability was assessed computing the *DSC* between:

- the UNet prediction masks and the respective ground-truth cartilage for the *Test sets 1-5* (*Table 2*); and
- the original ground-truths and the respective re-contoured cartilage *Volumes 1-5* (defined in *Intra-observer variability tests*).

The UNet prediction time for the US images in *Test sets 1-5* was also computed.

Evaluation 2

As in *Evaluation 1*, the standard *DSC* was measured. This iteration evaluated the UNet and the intra-observer variability on a subset of US images, and on the region of interest (ROI) in each of these images where clear hyperechoic cartilage boundaries were present (hereinafter referred as to *Tests 1-5*$_{sub}$ and *Tests 1-5*$_{subROI}$ respectively). The expert selected these images during a visual inspection of all the 2D US slices within the US volumes (*Volumes 1-5)* in the *Intra-observer variability tests.*

An example of US image selected for this study is shown in *Figure 6b* and it is compared to an example of an US image that was discarded, where the cartilage was present but not clearly delimited (*Figure 6a).* The region of interest in each of the images was manually selected, including only the area where both the inferior and superior hyper-echoic cartilage boundaries were clearly visible (*Figure 6b*).

The DSC was computed between:

- the UNet prediction masks and the respective original ground-truths;
- the UNet prediction masks and the respective re-contoured ground-truths;
- the original ground-truths and the respective re-contoured masks for the intra-observer variability assessment.

This evaluation aims at comparing the UNet and the expert performance while boundary uncertainty is reduced as much as possible. Moreover, measuring the performance over the reported image selection allowed the comparison of the intra-observer variability with other studies aiming at cartilage segmentation. In these studies, only images where clear hyper-echoic boundaries delimited the cartilage were considered (Faisal et al. 2018).

Evaluation 3

As per the last evaluation method, the developed DSC_{UB} (*Evaluation metrics*) was computed for*:*

- the UNet prediction masks and the respective ground-truths for the *Test sets 1-5* (*Table 2*);

- the original ground-truths and the respective re-contoured masks *Volumes 1-5* (*Intra-observer variability tests*).

The DSC was computed by defining the dimensions of the uncertainty margin based on intra-observer variability. The choice was made to obtain a conservative measure for the uncertainty margin by selecting only the images in *Intra-observer variability tests* where the cartilage boundaries were well defined (subset of US images in *Evaluation 2*). The original ground-truth mask were then overlapped with the corresponding re-contoured images in the subset considered and the pair-wise discrepancies were computed. The mean value determined suggested a uniform margin of +/- 0.4 mm that has been applied to each of the ground-truths considered in this assessment.

Results

Evaluation 1

The UNet and the expert performance related to *Evaluation 1* are reported in *Table 4*. For the UNet, the mean *DSC* over each test set (*Test sets 1-5)* varied between 0.65 - 0.71, with an overall mean value of 0.68. The algorithm performance on the dataset where femoral cartilage pathology was detected (*Test 4* in *Table 4*) was comparable to the results obtained in the other test sets. Figure 7 shows an example of an US image where the femoral cartilage defect was present and the expert and the algorithm cartilage segmentation has been superimposed on the US image represented in red and green, respectively. With regards to processing time, the UNet was able to segment a 2D US image in about 0.008 seconds, and hence, it can outline approximately 125 2D US images per second.

The expert scored a mean *DSC* over the volumes in each test set (*Volumes 1-5*) between 0 - 0.77 and an overall mean *DSC* of 0.64. The case where the expert scored 0 *DSC* (*Test 1* in *Table 4*) represents a peculiar case where none of the 2D US images of a volume contoured the first time were re-contoured when the volume was examined the second time. In this US volume (*Volume 1*), only the very extreme edge of the femoral cartilage was imaged. In fact, during the first contouring phase,

the femoral cartilage was identified only in ten US slices out of 256, and on these images, the cartilage boundaries were not clearly visible. A representative example of an annotated US slice is shown in *Figure 8*. Furthermore, during the first contouring phase, the annotator outlined the whole 4D sequence in chronological order from the first volume frame until the last frame, and that has possibly assisted the identification of cartilage position in this last volume frame (*Volume 1*).

Evaluation 2

Table 5 describes the UNet and the expert performance evaluated on a subset of 54 images selected from *Volumes 1-5* where the cartilage had clear hyperechoic boundaries (*Tests 1-5*$_{sub}$) and on the actual region of each of these images where this condition was true (*Tests 1-5*$_{subROI}$) (*Evaluation 2*). No images with well-defined cartilage boundary were found for *Volume 1* (*Result Evaluation 1* and *Figure 8*). A total of 26, 19, 6 and 4 images were selected for *Volumes 2-5* respectively. The first column of *Figure 9* shows examples of the images and the image regions selected (outlined by a yellow box) for this part of the study. Viewing the figure from top to bottom, the selected US images correspond to *Volumes 2-5*, respectively. For each US image in the figure, the segmentations produced by the UNet, by the expert during the ground-truth creation and the intra-observer test are shown in green, red and blue respectively.

Considering that this assessment evaluated the UNet and the expert performance precisely in the same images/image regions, the UNet could be compared to both the original ground-truth and the images that the expert re-contoured (referred as to *UNet*^{1} *and UNet*^{2} in *Table 5*). When compared to the average *DSC* values in *Results Evaluation 1*, the *DSC* increased significantly both for the UNet and the expert, reaching higher values when evaluated on the selected image regions (*Table 5*). The overall UNet *DSC* was around 1-2% higher than the expert *DSC* for both the images and the image regions selected.

Evaluation 3

Table 6 reports the results corresponding to the UNet and the expert performance measured with the $DSC_{UB.}$ The mean DSC_{UB} for the UNet ranged between 0.85-0.91, with an overall mean value of 0.87. As in *Results, Evaluation 1* the algorithm performance on the dataset with femoral cartilage pathology (*Test 4* in *Table 6*) was comparable to the results obtained in the other test sets.

Contrasting to the previous evaluation, the mean DSC_{UB} achieved by the expert was between 0.51-0.91, with an overall mean of 0.78.

Discussion

This paper has presented the first attempt of using a CNN (UNet) for femoral cartilage segmentation applied to dynamic, volumetric US for robotic knee arthroscopy. This work aims to localise the femoral cartilage using intra-operative US imaging to avoid collision/contact between the surgical tools and anatomical structures. Accordingly, the priority was to detect the cartilage whenever possible with the highest accuracy, even when the tissue boundary sharpness in the US images was not optimal. This inherent lack of information in the images resulted in a high intra-observer variability. In many cases, the expert was not able to produce precise and consistent contours (*Results, Evaluation 1*). In some critical cases, the overlap between the contours outlined in two different sessions was zero.

A strong correlation between the expert performance and the uncertainty at the cartilage boundary in the US images is evident as it can be demonstrated from the *DSC* values for those images/image regions with clear hyperechoic boundaries (*Results, Evaluation 2)*. In those cases, when sufficient information is present in the image, the expert was consistent and performed comparably to clinical standards (Faisal et al. 2018a) (*Results, Evaluation 2*).

This result suggests that the intra-operator variability may be used to obtain an implicit measure of the uncertainty margin surrounding the target and could provide an estimate for the dimensions of the cartilage boundary region where the expert was less consistent. The intention was to calculate an approximate margin based on those images where boundary uncertainty was minimal

and apply this conservative margin to all the ground-truths. This choice is justified by the limited dataset used for the intra-operator study (*Intra-observer variability tests*). Alternatively, a local variable margin expansion could be assigned by the expert to the different ground-truth contour regions based on a clinical evaluation of the uncertainty in that specific area.

The concept of uncertainty margin introduced in this work was used to generate a more accurate metric (the DSC_{BU}) to evaluate the algorithm and the expert performance (*Evaluation metrics*). The DSC_{BU} might apply to all those cases in medical imaging where the boundaries of the target are not defined well enough to be represented by a contour line (Hindi et al. 2013; Shrimali et al. 2009) and may be a useful application as this is particularly pervasive issue in US imaging. With the metric evaluated, the distance between the ground-truth and the contour generated by the algorithm includes a margin expansion for the contour in the regions which are more prone to variability because they are less identifiable in the US images.

Segmentation uncertainty has recently become a topic of interest in the field of DL, where different CNNs have been developed to predict a probability map of the pixels within a segmentation (Hall et al. 2018; Isobe and Arai; Kendall and Roberto Cipolla 2016; Nair et al. 2018). It is the intention of the authors to investigate how the new metric compares to the standard *DSC* for these types of algorithms.

The results of the evaluation tests indicate that the UNet performed at least as well as an experienced clinician, and has solid potential for segmenting the femoral cartilage under simulated surgical conditions. For some of the test sets considered in the *DSC* and the DSC_{BU} evaluations (*Results, Evaluation 1* and *Results, Evaluation 3)*, significant discrepancies were found between the UNet and the expert performance. However, in those cases, while the UNet was evaluated on the entire test sets, the expert performance was assessed only on one volume in each test set, and thus the discrepancy in the results might be due to the particular choice of the test volumes. It should be noted that when the UNet and the expert were compared using the same images with clear cartilage boundary, no relevant difference was found between the CNN and the expert (*Results, Evaluation 2*).

However, there was a substantial difference in the information given to the surgeon and the CNN to produce the cartilage segmentation. The surgeon could visualise the whole US volume and the previous/consequent contoured image slices within the volume while creating the ground-truth masks (*Labels generation*). The UNet instead processes a 2D image at the time and does not have any additional information about the rest of the volume, making the cartilage identification task even more challenging. In a future study, it would be interesting to compare the UNet performance with a 3D CNN, e.g. (Çiçek et al. 2016; Milletari et al. 2017). For the final application a 3D approach would be possibly preferred since volumetric US will be used for guidance.

This study has demonstrated that the UNet is currently able to segment 125 2D US images per second which corresponds on average to the number of 2D US slices to be contoured in each US volume. Thus, running multiple UNets in parallel one could potentially reach real-time volumes segmentation (e.g. 30 volumes per second). Nevertheless, the full volume rate of the US probe used in this study is limited to one full volume per second and to the best of our knowledge there are presently no existing volumetric US probes in the correct frequency range and with the correct imaging characteristics that allow for a faster acquisition. In the next future, these requirements may be achieved by emerging US technologies, e.g. xMATRIX transducers.

One of the main limitations of this work is the restricted range of motion of the knee joint (0 to 30 degrees flexion angles) and the associated possible surgical scenarios, as opposed to the real surgical process where the leg may be flexed to larger angles ((McKeon et al. 2009). With this noted, this research has considered the lesser angles of flexion which result in the greatest risk of operative damage to the femoral cartilage due to larger areas of the structure being exposed to the surgical tools. As the knee flexion angle increases, the medial-lateral femoral condyles shift towards the posterior direction due to the patella-femur roto-translation, and the femoral cartilage slides beyond the patella, moving away from the surgical site (Paczesny and Kruczyński 2011). Future studies will required to assess the femoral cartilage in the US images for flexion angles greater than 30 degrees and evaluate the related possible risk of damage. Similar types of scans as the ones reported in this study (i.e.

where the probe is translated at a fixed knee angle or the knee extended) could be applied to the whole range of knee angles that may occur in the operating theatre. Additionally, different US settings should be considered to ensure the algorithm robustness to these types of image variations.

Another important topic of investigation is the algorithm response to US images showing femoral cartilage pathologies. In this paper, it was demonstrated that the algorithm is effective with considerable variability of cartilage appearance, even when tested on US images examples of cartilage defects. Femoral cartilage degeneration is characterised by different stages of severity, from minor damage, e.g. focal defects (i.e. small lesions at the cartilage surface) to severe, e.g. full thickness degeneration (i.e. part of the cartilage is missing) (Aisen et al. 2014; Möller et al. 2008; Ohashi et al. 2012; Paczesny and Kruczyński 2011; Qvistgaard et al. 2006). Testing the algorithm extensively on unhealthy cartilage is an essential step before clinical translation. This could be achieved by including examples of different stages of femoral cartilage pathology into the training dataset. Such a training set would allow the deep learning algorithm to recognize the appearance of pathologic cartilage and consequently directly interpret this source of variability.

Another limitation to this investigation, is the relatively small number of volunteers. This choice is justified by the large number of annotations required (>18000) to assess the different surgical scenarios. A possible efficient solution to create a larger dataset may be achieved by utilising the trained CNN to generate the cartilage segmentation for subjects' US images and adding the segmentations compatible to clinical standards to the CNN training set. This process could be repeated multiple times, resulting in re-training of the CNN to increment the dataset each iteration and potentially improve the algorithm performance, hence reducing the need for contouring additional images.

This proof of concept report investigates the clinical applicability of US imaging in creating a complete map of the knee tissues which would enable US-based navigation of a potentially autonomous robotic knee arthroscopy system (Wu et al. 2019). Future focus will be the addition of magnetic resonance imaging (MRI) of the volunteers' knees which will then be co-registered to the

respective US volumes to provide supplementary anatomical information. This process will be paramount when multiple knee structures need to be identified simultaneously from the US volumes, as presently no guidelines exist. The UNet could be subsequently applied as a multi-label segmentation algorithm to map the surgical site for this type of application automatically.

Conclusion

The reported findings represent a first attempt of utilising a UNet to segment femoral cartilage on images extracted from 3D US dynamic acquisitions, collected under simulated surgical conditions. The sharpness of the cartilage boundary was variable between the considered images and frequently not clearly defined. This was addressed by assessing the uncertainty at the cartilage boundary through an intra-observer study and introducing a revised version of the Dice similarity coefficient (the DSC_{BU}) that can account for this uncertainty. The CNN performance is comparable to an expert with DSC_{BU} of 0.87, indicating its significant potential to identify the femoral cartilage during robotic knee arthroscopy. These findings represent a component of a larger project investigating the feasibility of using US imaging for robotic knee arthroscopy.

Acknowledgements

This work is part of the Australia-India strategic research fund AISRF53820 (Intelligent Robotic Imaging System for keyhole surgeries). The computational resources and services used in this project were provided by the HPC and Research Support Group, Queensland University of Technology, Brisbane, Australia. G.C acknowledges the support by Australian Research Council through grant DP180103232.

References

Aisen AM, McCune WJ, MacGuire A, Carson PL, Silver TM, Jafri SZ, Martel W. Sonographic evaluation of the cartilage of the knee. Radiology 2014.

Aka NORAB, Eenstra SIL, Alsum. Random forest-based bone segmentation in ultrasound. Ultrasound Med Biol 2017;43:2426–2437.

Alves TI, Girish G, Brigido MK, Jacobson JA. US of the Knee: Scanning Techniques, Pitfalls, and Pathologic Conditions. RadioGraphics 2016;36:1759–1775.

Antico M, Sasazawa F, Wu L, Jaiprakash A, Roberts J, Crawford R, Pandey AK, Fontanarosa D. Ultrasound guidance in minimally invasive robotic procedures. Med Image Anal 2019;54:149–167.

Cianca J, John J, Pandit S, Chiou-Tan FY. Musculoskeletal ultrasound imaging of the recently described anterolateral ligament of the knee. Am J Phys Med Rehabil 2014;93:186.

Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-net: Learning dense volumetric segmentation from sparse annotation. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics) 2016;9901 LNCS:424–432.

Curl WW, Krome J, Gordon ES, Rushing J, Smith BP, Poehling GG. Cartilage injuries: A review of 31,516 knee arthroscopies. Arthroscopy. 1997. pp. 456–460.

Dice LR. Measures of the Amount of Ecologic Association Between Species. Ecology 2006;26:297–302.

Faisal A, Ng S, Goh S, Lai KW. Knee Cartilage Ultrasound Image Segmentation Using Locally Statistical Level Set Method. 2nd Int Conf Innov Biomed Eng Life Sci IFMBE Proc 2018a;:275–281.

Faisal, Ng S-C, Goh S-L, Lai KW. Knee cartilage ultrasound image segmentation using locally statistical level set method. IFMBE Proc 2018b.

Faisal, Ng SC, Goh SL, Lai KW. Knee cartilage segmentation and thickness computation from ultrasound images. Med Biol Eng Comput 2018c;56:657–669.

Grzelak P, Podgórski MT, Stefańczyk L, Domżalski M. Ultrasonographic test for complete anterior

cruciate ligament injury. Indian J Orthop 2016;49:143–9.

Guerrero J, Salcudean SE, McEwen JA, Masri BA, Nicolaou S. Real-time vessel segmentation and tracking for ultrasound imaging applications. IEEE Trans Med Imaging 2007;26:1079–1090.

Hall D, Dayoub F, Skinner J, Corke P, Carneiro G, Sünderhauf N. Probability-based Detection Quality (PDQ): A Probabilistic Approach to Detection Evaluation. arXiv:181110800 2018;

Hindi A, Peterson C, Barr RG. Artifacts in diagnostic ultrasound. Reports Med Imaging 2013;6:29–48.

Huang Q, Zhang F, Li X. Machine Learning in Ultrasound Computer-Aided Diagnostic Systems : A Survey. Biomed Res Int Hindawi, 2018.

Ioffe, Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. arXiv:150203167 2015.

Isobe S, Arai S. Inference with model uncertainty on indoor scene for semantic segmentation. 2017 IEEE Glob Conf Signal Inf Process Glob 2017 - Proc.

Jaiprakash A, O'Callaghan WB, Whitehouse SL, Pandey A, Wu L, Roberts J, Crawford RW. Orthopaedic surgeon attitudes towards current limitations and the potential for robotic and technological innovation in arthroscopic surgery. J Orthop Surg 2017;25.

Jaumard-Hakoun A, Xu K, Roussel-Ragot P, Dreyfus G, Denby B. Tongue contour extraction from ultrasound images based on deep neural network. Proc 18th Int Congr Phonetic Sci (ICPhS 2015) 2016.

Kendall A, Roberto Cipolla. Bayesian SegNet: Model Uncertainty in Deep Convolutional Encoder-Decoder Architectures for Scene Understanding. arXiv:151102680 2016.

Kingma DP, Ba JL. Adam: A method for stochastic gradient descent. ICLR Int Conf Learn Represent 2015.

Kirkpatrick S, Gelatt CD, Vecchi MP. Optimization by simulated annealing. Science (80-) 1983;220:671–680.

Kowal J, Amstutz C, Langlotz F, Talib H, Ballester MG. Automated bone contour detection in

ultrasound B-mode images forminimally invasive registration in computer-assisted surgery – an in vitro evaluation. Int J Med Robot Comput Assist Surg 2007;3:341–348.

Lueders DR, Smith J, Sellon JL. Ultrasound-Guided Knee Procedures. Phys Med Rehabil Clin N Am Elsevier Inc, 2016;27:631–648.

McKeon BP, Bono J V, Richmond JC. Knee Arthroscopy. 2009.

Milletari F, Ahmadi SA, Kroll C, Plate A, Rozanski V, Maiostre J, Levin J, Dietrich O, Ertl-Wagner B, Bötzel K, Navab N. Hough-CNN: Deep learning for segmentation of deep brain regions in MRI and ultrasound. Comput Vis Image Underst 2017;164:92–102.

Milletari F, Navab N, Ahmadi SA. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. Proc - 2016 4th Int Conf 3D Vision, 3DV 2016 2016.

Möller I, Bong D, Naredo E, Filippucci E, Carrasco I, Moragues C, Iagnocco a. Ultrasound in the study and monitoring of osteoarthritis. Osteoarthritis Cartilage 2008;16 Suppl 3:S4-7.

Nair T, Precup D, Arnold DL, Arbel T. Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics) 2018.

Ohashi S, Ohnishi I, Matsumoto T, Bessho M, Matsuyama J, Tobita K, Kaneko M, Nakamura K. Measurement of Articular Cartilage Thickness Using a Three-Dimensional Image Reconstructed from B-Mode Ultrasonography Mechanical Scans Feasibility Study byComparison with MRI-Derived Data. Ultrasound Med Biol 2012.

Paczesny Ł, Kruczyński J. Ultrasound of the Knee. Semin Ultrasound, CT MRI 2011;32:114–124.

Paszke A, Chanan G, Lin Z, Gross S, Yang E, Antiga L, Devito Z. Automatic differentiation in PyTorch. 31st Conf Neural Inf Process Syst (NIPS 2017) 2017;1–4.

Price AJ, Erturan G, Akhtar K, Judge A, Alvand A, Rees JL. Evidence-based surgical training in Orthopaedics: How many arthroscopies of the knee are needed to achieve consultant level performance? Bone Jt J 2015;97-B:1309–1315.

Qvistgaard E, Torp-Pedersen S, Christensen R, Bliddal H. Reproducibility and inter-reader

agreement of a scoring system for ultrasound evaluation of hip osteoarthritis. Ann Rheum Dis 2006;65:1613–1619.

Ravishankar H, Prabhu SM, Vaidya V, Singhal N. Hybrid approach for automatic segmentation of fetal abdomen from ultrasound images using deep learning. 2016 IEEE 13th Int Symp Biomed Imaging 2016. pp. 779–782.

Razek A, Fouda NS, Elmetwaley N, Elbogdady E. Sonography of the knee joint. J Ultrasound 2009;12:53–60.

Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. 2015;arXiv:1505:1–8.

Shrimali V, Anand R, Kumar V. Current Trends in Segmentation of Medical Ultrasound B-mode Images: A Review. IETE Tech Rev 2009;26:8–17.

Smistad E, Løvstakken L. Vessel detection in ultrasound images using deep convolutional neural networks. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). 2016. pp. 30–38.

Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. J Mach Learn Res 2004;1929−1958.

Vlad VM. Pitfalls in Musculoskeletal Ultrasound. Musculoskelet Ultrason Rheum Dis - Springer Int Publ 2015;21–30.

Wu L, Jaiprakash A, Pandey AK, Fontanarosa D, Jonmohamadi Y, Antico M, Strydom M, Razjigaev A, Sasazawa F, Roberts J, Crawford R. Robotic and Image-Guided Knee Arthroscopy. Elsevier's Handb Robot Image-Guided Surgery 2019. p. (in press).

Zarins B, Rowe CR, Harris BA, Watkins MP. Rotational motion of the knee*. Am J Sports Med 1983;11:152–156.

Figure Captions List

Figure 1: Femoral cartilage US scan (Faisal et al. 2018b). The cartilage is visualised in the US image as an anechoic (dark) layer, attached to the femoral condyles, with highly hyperechoic (very bright) boundaries.

Figure 2: US probe positioning. On the left: The volunteer knee joint is scanned in water with the US probe placed on the patellar tendon. On the right: schematic US probe positioning representation showing the positions of reference structures relative to the probe.

Figure 3: GUI for cartilage contouring. The interface shows the GUI components and an example of an US volume with the cartilage outlined by the expert in the sagittal plane (dark green contour). As an additional source of information, the annotator could also visualise the projection of the contour drawn on the subsequent slice (light green). The contours generated along the sagittal plane are shown in the axial and coronal projections and the yellow crosshair indicates a common point along the three planes.

Figure 4: Graphical visualisation of the UNet architecture. The blue blocks represent intermediate feature maps and the values above them indicate their depth. Coloured arrows show the operations that are applied to input data and feature maps. The blocks on the left of the image forms the encoding branch; while the right blocks define the decoding branch.

Figure 5: Schematic representation of the segmentations involved in the computation of the Dice similarity coefficient with boundary uncertainty (DSC_{BU}). In Figure a, the standard ground-truth M_{GT} is outlined in black; the uncertainty margin UM_{GT} is represented as a red mask bounding the standard ground-truth, indicating the area where the tissue boundaries can be located. The internal ground-truth I_{GT} (the part of the ground-truth delimited by the internal boundary of the UM_{GT}) is shown in blue and represents the region that undoubtedly belongs to the target tissue. Masks enclosing the I_{GT} and having boundaries within the UM_{GT} can be considered as acceptable ground-truths. Some examples are outlined with dashed lines in different colours (green, yellow and white). Figure b shows an example of prediction mask M_P (outlined with a yellow dashed line) and the ground-truth mask with boundary uncertainty $M_{GT_{BU}}$ (dashed blue mask) defined as the sum of the I_{GT} and the intersection between the prediction M_P and the uncertainty margin mask UM_{GT}. The $M_{GT_{BU}}$ is selected as the ground-truth for the target as it contains the I_{GT} and has boundaries in the UM_{GT}. In this way the pixels of the M_P located in the uncertainty margin UM_{GT} will not be penalised in the dice computation.

Figure 6: Examples of US images without/with clear femoral cartilage boundaries. In Figure a, an US image discarded from *Evaluation 2* where the cartilage was present but the exact location of the boundaries is unclear. The area where the cartilage is located is encircled in yellow. In Figure b, an example of US image selected for *Evaluation 2* and the region of the same image selected (in red) where both superior and inferior hyperechoic cartilage boundaries are visible (highlighted in green).

Figure 7: US image example with femoral cartilage partial-thickness degeneration (yellow). The green and red segmentations produced by the UNet and the expert, respectively.

Figure 8: Example of a 2D US annotated slice from *Volume 1*. The figure shows a 2D US image with the corresponding femoral cartilage contour in yellow. *Volume 1* is the last frame of a 4D sequence where the probe was translated towards the tibia (highlighted in green) *(Translation 0* scanning type,

Image data acquisition). For this reason, the 2D US image shows mainly the tibia and only a small edge of the lateral femoral cartilage.

Figure 9: Examples of the images/image regions selected for *Evaluation 2* and resulting segmentations. The first column of the figure shows examples of the images and the image regions selected (outlined by a yellow box) for this part of the study. The UNet and the expert performance are evaluated either considering the whole images (*Tests 1-5*$_{sub}$ in *Table 5*) or only the yellow region highlighted (*Tests 1-5*$_{subROI}$ in *Table 5*).

Tables

Table 1: 4D US acquired from 6 volunteers. Columns 1-6 describe the volunteer information (volunteers ID, sex, age, weight and height, leg scanned and femoral cartilage pathologies). Column 7 reports the number of 4D sequences acquired. Columns 8 reports the three types of scans performed. The number of 3D volumes (extracted from the 4D US sequences) per type of scan (for both legs) and the total number of 3D volumes per volunteer are shown in Columns 9-10 respectively.

Volunteer ID	Sex	Age	Weight [kg] Height [cm]	Leg	Femoral cartilage pathologies	#4D sequences	Scan Type	#3D volumes per scan type	#3D volumes
1	M	34	60 171	L, R		6	Extension 30 Translation 0 Translation 30	10 10 8	28
2	F	34	64 170	L,R	-	6	Extension 30 Translation 0 Translation 30	9 8 6	23
3	M	31	71 185	L,R	-	6	Extension 30 Translation 0 Translation 30	9 8 11	28
4	M	44	80 183	L,R	Partial thickness degeneration in both legs	6	Extension 30 Translation 0 Translation 30	6 10 8	24
5	M	34	78 180	L,R	-	6	Extension 30 Translation 0 Translation 30	12 8 9	29
6	F	20	43 153	L,R	-	5	Extension 30 Translation 0 Translation 30	3 10 9	22

Table 2: Datasets (*Dataset 1-5*) for the UNet cross-validation. The entire dataset is split in *Training* and *Test sets* (Columns 3-4), where the *Dataset n* is the data partition having in the *Test set n* that contains 2D labelled US images for volunteer *n*.

Dataset IDs	Scan Type	Train set		Test set	
		#2D labelled images per scan type	Volunteer IDs	#2D labelled images per scan type	Volunteer IDs
1	Extension 30 Translation 0 Translation 30	4553 5388 5217	2, 3, 4, 5, 6	1197 1192 730	1
2	Extension 30 Translation 0 Translation 30	4141 5095 5069	1, 3, 4, 5, 6	1609 1385 878	2
3	Extension 30 Translation 0 Translation 30	4712 5665 4398	1, 2, 4, 5, 6	1038 815 1549	3

4[b]	Extension 30	5015	1, 2, 3, 5, 6	735	4
	Translation 0	5383		1097	
	Translation 30	4535		1412	
5	Extension 30	4676	1, 2, 3, 4, 6	1074	5
	Translation 0	5494		1086	
	Translation 30	5450		497	

[b] Volunteer with femoral cartilage pathology

Table 3: Description of the volumes re-contoured for the Intra-observer test. *Volumes 1-5* are selected within the *Test sets 1-5* respectively. Columns 2-4 report the characteristics of each volume: the leg scanned; the scan type and the volume frame in the 4D sequence (e.g. volume frame 5 is in the fifth volume recorded in chronological order within the 4D sequence).

Volumes	**Leg**	**Scan Type**	**Volume Frame**
Volume 1	L	Translation 0	5
Volume 2	L	Extension 30	5
Volume 3	R	Translation 0	2
Volume 4	L	Translation 30	1
Volume 5	L	Extension 30	4

Table 4: Standard *DSC* achieved by the UNet and by the expert. Columns 2-5 report the mean *DSC* and the minimum/maximum *DSC* values within the test sets; columns 6 and 7 show the mean *DSC* computed excluding the zero *DSC* predictions (named DSC_0) and the percentage in the test sets where this condition was true. In the last row of the table, the overall performance of the UNet and the expert was computed as the mean *DSC* and mean DSC_0 over the images considered in the test sets.

# Test[a]	**Mean DSC**		**Min/Max DSC**		**Mean DSC_0 (# 0 predictions [%])**	
	UNet	**Intra-observer**	**UNet**	**Intra-observer***	**UNet**	**Intra-observer**
Test 1	0.65	0	0 / 0.95	-	0.67 (6%)	- (100%)
Test 2	0.68	0.77	0 / 0.94	0 / 0.95	0.73 (6%)	0.80 (2%)
Test 3	0.69	0.73	0 /0.95	0 / 0.92	0.72 (4%)	0.77 (2%)
Test 4[b]	0.68	0.40	0 / 0.94	0 / 0.92	0.71 (4%)	0.68 (41%)
Test 5	0.71	0.70	0 / 0.94	0 / 0.92	0.73 (2%)	0.78 (12%)
Overall	**0.68**	**0.64**	**-**	**-**	**0.72 (4%)**	**0.77 (16%)**

[a] *Tests 1-5* correspond to the evaluation of the respective *Test sets 1-5* for the UNet and *Volumes 1-5* for the expert (*Evaluation 1*)
[b] Volunteer with femoral cartilage pathology

Table 5: Standard *DSC* achieved by the UNet and by the expert, computed on a subset of images/image regions with clear hyperechoic cartilage boundaries. *Tests 1-5*$_{sub}$ and *Tests 1-5*$_{subROI}$ correspond to the evaluation of the images and image regions selected respectively from *Volumes 1-*

5 (s*Evaluation 2*). Columns 2-7 report the mean *DSC* and the minimum/ maximum *DSC* values within the test sets. In the last row of the table, the overall performance of the UNet and the expert was computed as the mean *DSC* over all the images/image regions considered in the test sets.

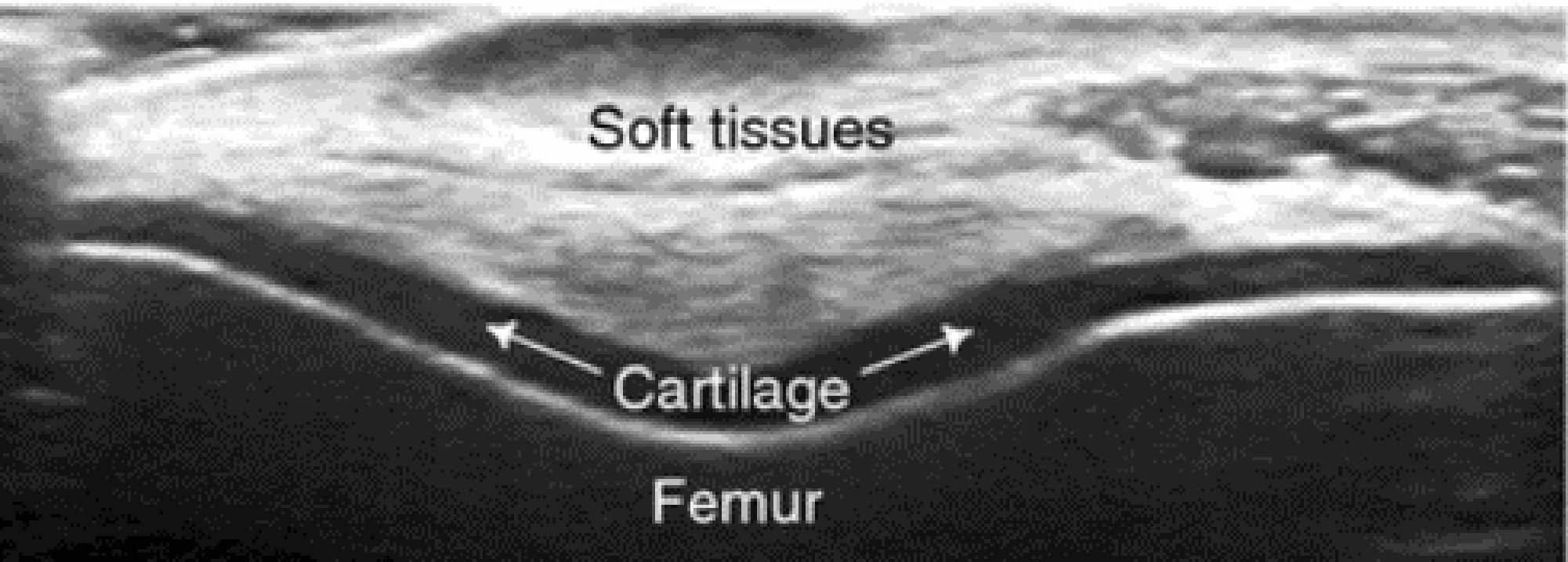
# Test $_{sub}$	**Mean DSC**			**Min/Max DSC**		
	UNet[1a]	**UNet[2a]**	**Intra-observer**	**UNet[1a]**	**UNet[2a]**	**Intra-observer**
Test 1_{sub}	-	-	-	-	-	-
Test 1_{subROI}	-	-	-	-	-	-
Test 2_{sub}	0.84	0.81	0.84	0.70/0.89	0.75/0.87	0.74/0.92
Test 2_{subROI}	0.90	0.85	0.88	0.81/0.95	0.79/0.91	0.77/0.96
Test 2_{sub}	0.85	0.89	0.83	0.68/0.92	0.85/0.93	0.75/0.91
Test 2_{subROI}	0.90	0.92	0.88	0.79/0.96	0.89/0.96	0.81/0.95
Test 3_{sub}	0.78	0.90	0.78	0.73/0.81	0.88/0.91	0.70/0.82
Test 3_{subROI}	0.86	0.91	0.87	0.83/0.88	0.88/0.94	0.86/0.90
Test 4_{sub}[b]	0.84	0.81	0.77	0.79/0.89	0.78/0.87	0.69/0.86
Test 4_{subROI}[b]	0.84	0.81	0.77	0.79/0.90	0.78/0.88	0.69/0.86
Overall	**0.84** **0.89**	**0.85** **0.88**	**0.83** **0.87**	-	-	-

[a] *UNet[1]* corresponds to the UNet compared to the original ground-truths; *UNet[2]* corresponds to the UNet compared to the ground-truth images re-contoured by the expert for the intra-observer study
[b] Volunteer with femoral cartilage pathology

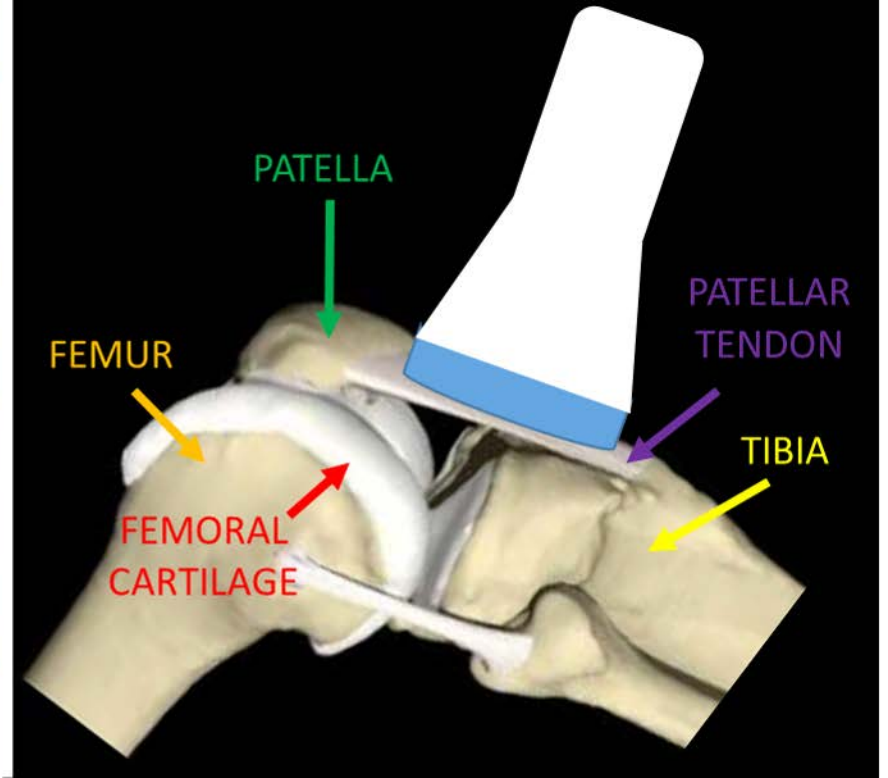
Table 6: DSC_{UB} achieved by the UNet and by the expert. Columns 2-5 report the mean DSC_{UB} and the minimum/maximum DSC_{UB} values within the test sets; columns 6 and 7 show the mean DSC_{UB} computed excluding the zero predictions (named DSC_{UB0}) and the percentage in the test sets where this condition was true. In the last row of the table, the overall performance of the UNet and the expert was computed as the mean DSC_{UB} and mean DSC_{UB0} over the images considered in the test sets.

# Test[a]	**DSC_{UB}**		**Min/Max DSC_{UB}**		**DSC_{UB0} (# 0 predictions [%])**	
	UNet	**Intra-observer**	**UNet**	**Intra-observer**	**UNet**	**Intra-observer**
Test 1	0.87	/	0/1	/	0.90 (6%)	- (100%)
Test 2	0.91	0.91	0/1	0/1	0.93 (6%)	0.94 (2%)
Test 3	0.86	0.88	0/1	0/1	0.91 (4%)	0.92 (2%)
Test 4[b]	0.86	0.51	0/1	0/1	0.91 (4%)	0.88 (41%)
Test 5	0.85	0.89	0/1	0/1	0.89 (2%)	0.99 (12%)
Overall	**0.87**	**0.78**	-	-	**0.90 (4%)**	**0.93 (16 %)**

[a] *Tests 1-5* correspond to the evaluation of the respective *Test sets 1-5* for the UNet and *Volumes 1-5* for the expert (*Evaluation 3*).
[b] Volunteer with femoral cartilage pathology.


Soft tissues
Cartilage
Femur


US PROBE

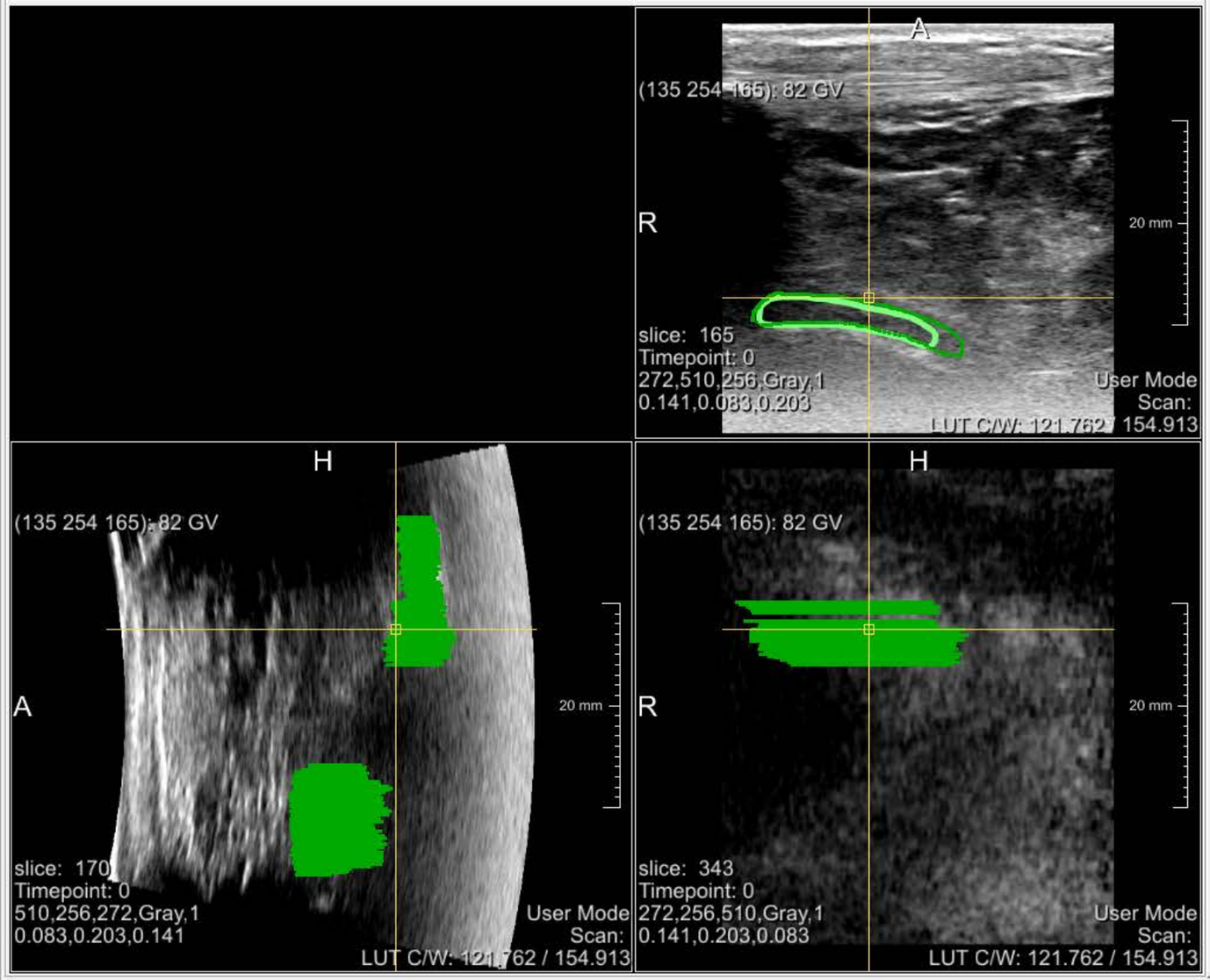

PATELLA
PATELLAR TENDON
FEMUR
TIBIA
FEMORAL CARTILAGE

Segmentations | Contour Viewer | Contour Image Viewer | Contour To Image Registration

Segmentations

Viewers

Cube Crosshair



File selection

File: _Quality_Tracking/FUMIO_TEST_11-01/Intra-observer_VR_11-01/Test1/IM_0008_frame2.mhd Browse...

Load CSO

File: h_Quality_Tracking/FUMIO_TEST_11-01/Intra-observer_VR_11-01/Test1/IM_0008_frame2.cso Browse...

Settings

☑ Ghost mode

CSOLIST

Commands

Remove All | Undo | Redo | Combine | Break

Parameters

CSO | Group | Default | Notifications | Selection

CSO [129]

Id	La
84	1:
85	1:
86	1:
87	1:

Delete Copy

Group [1]

	Id	Label	C
*	1	sagittal (7)	1,

New Delete Copy

Default CSO Parameters

CSO Common | Grp Common | Grp Visuals | Grp Misc

☐ Work directly on input CSOList | Copy Input CSOList | ☑ Enable Undo/Redo | Undo Stack Limit: 10

CSO_Axial | CSO_Sagittal | CSO_Coronal

Interpolators

Interpolate Axial Plane | Interpolate Sagittal Plane

CSO Ids: 1

Smooth Contours

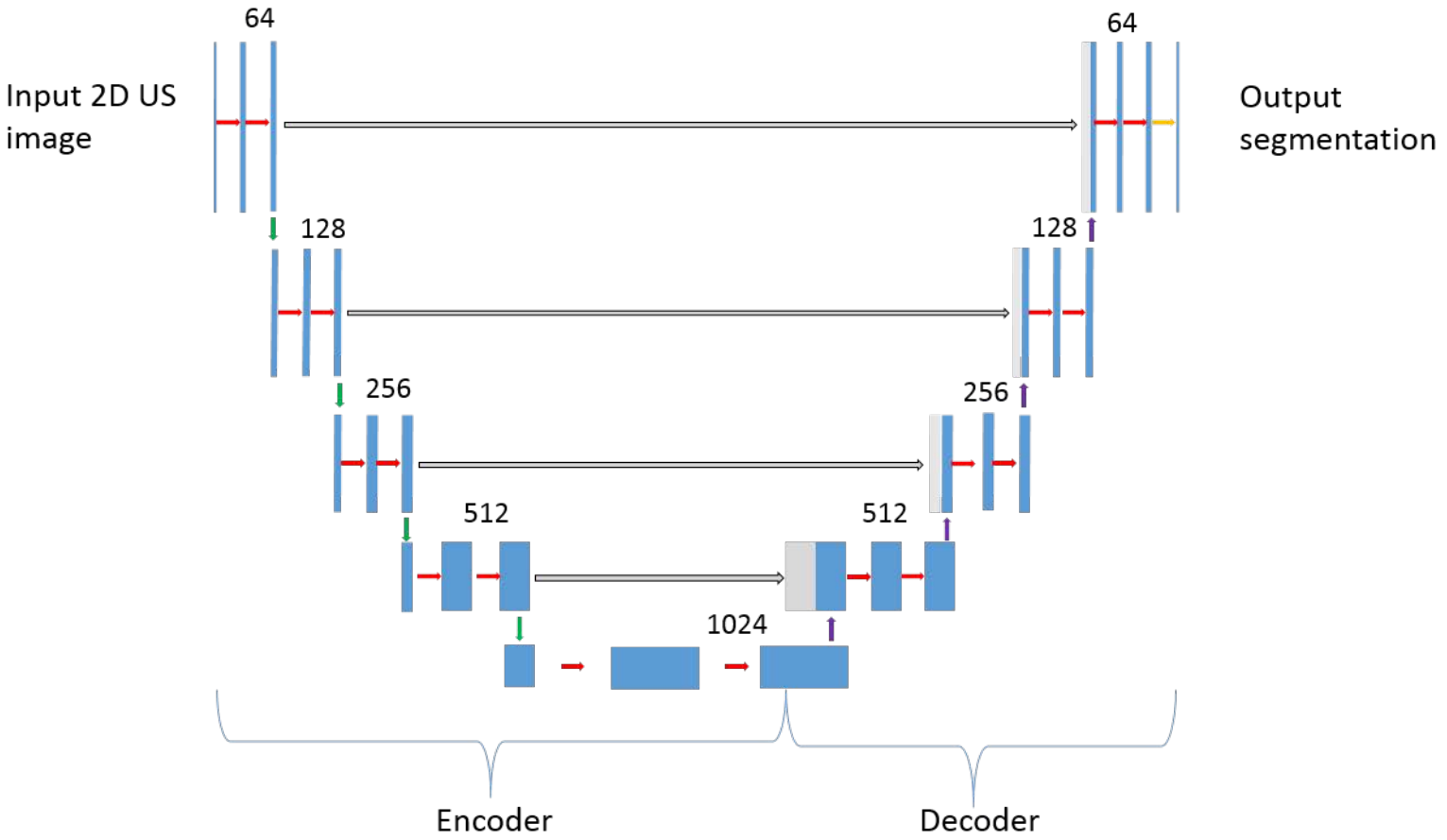
Save CSO

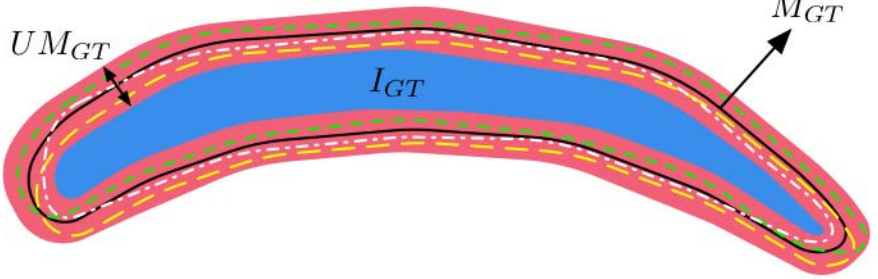
Update CSOList

File: g/FUMIO_TEST_11-01/Intra-observer_VR_11-01/Test6 - Copy/Unfinish_IM_00014_frame2.cso Browse...

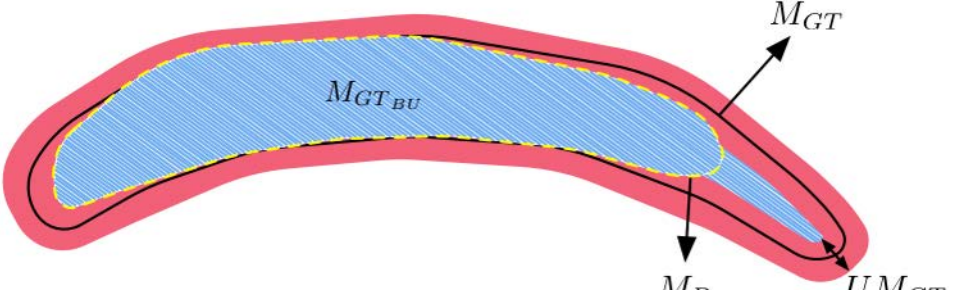
Convert CSO to Binary Image

Convert | File: 4.dcm Browse...

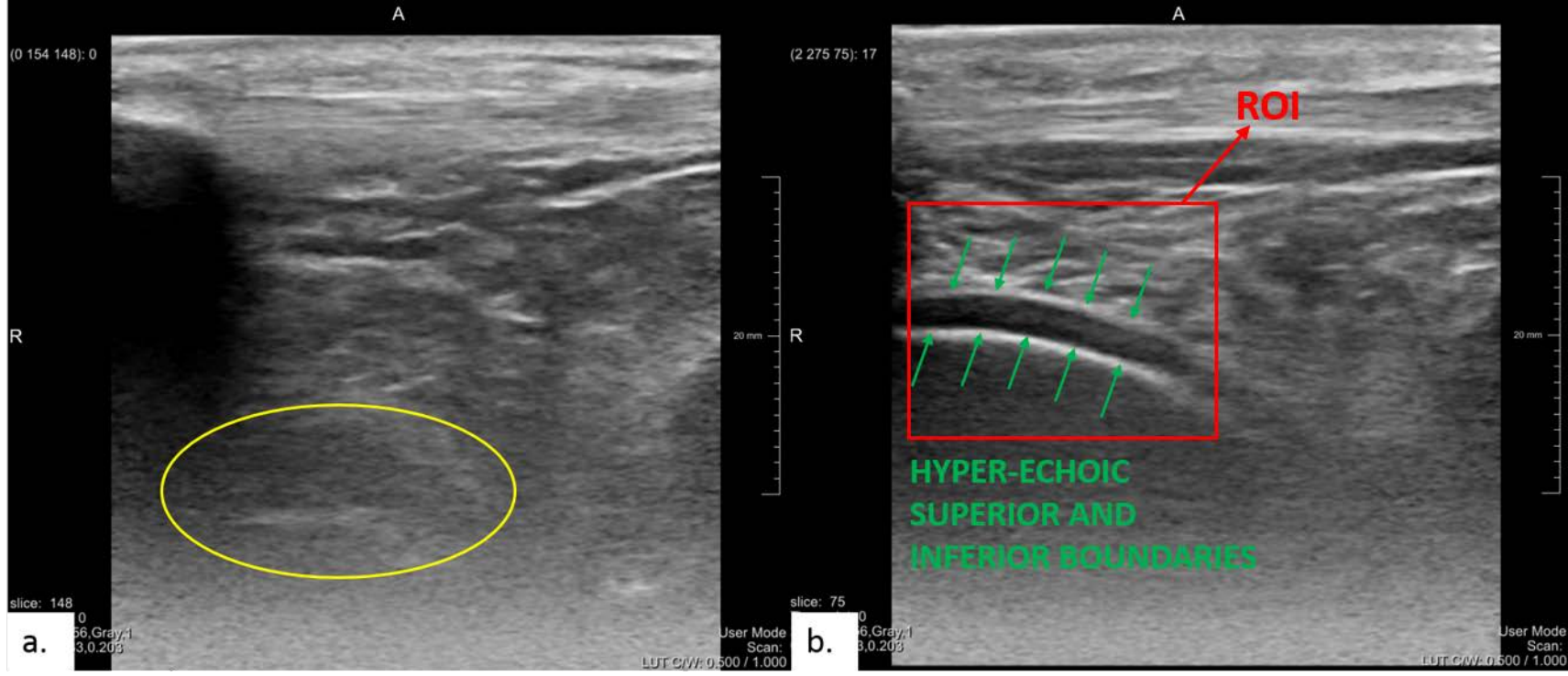

Input 2D US image
64
64
Output segmentation
128
128
256
256
512
512
1024
Encoder
Decoder
Conv 3x3 - Batch normalization - ReLu - Dropout
Max pool 2x2
Up-conv 2x2
Conv 1x1
Copy

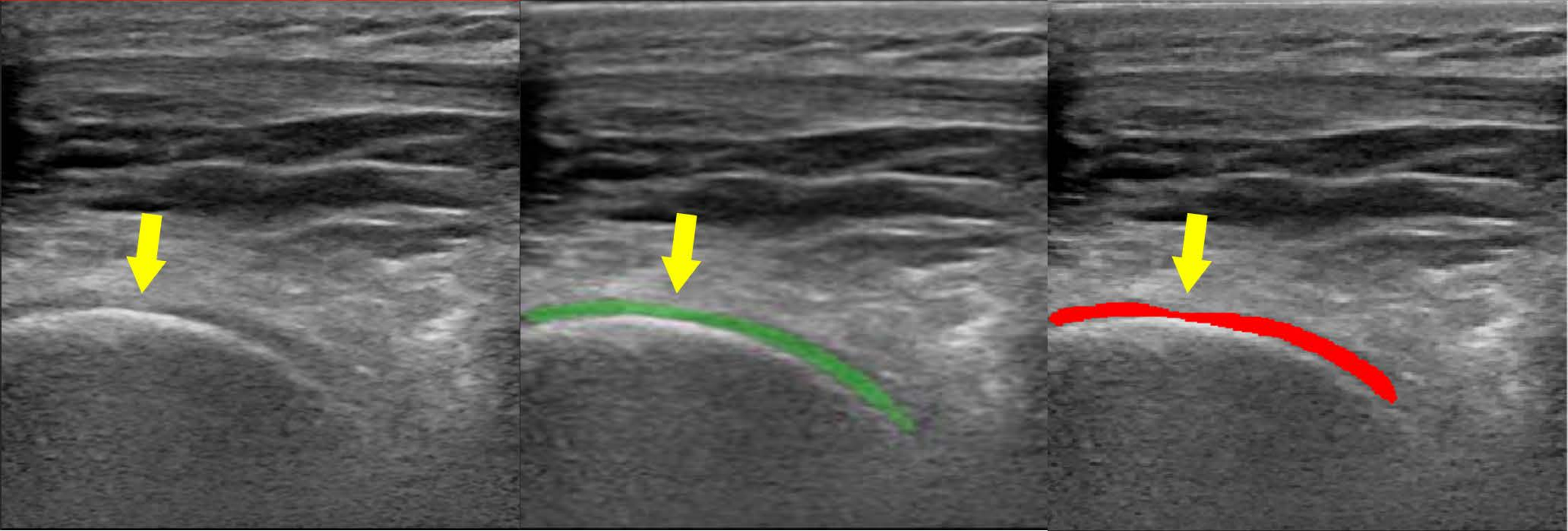


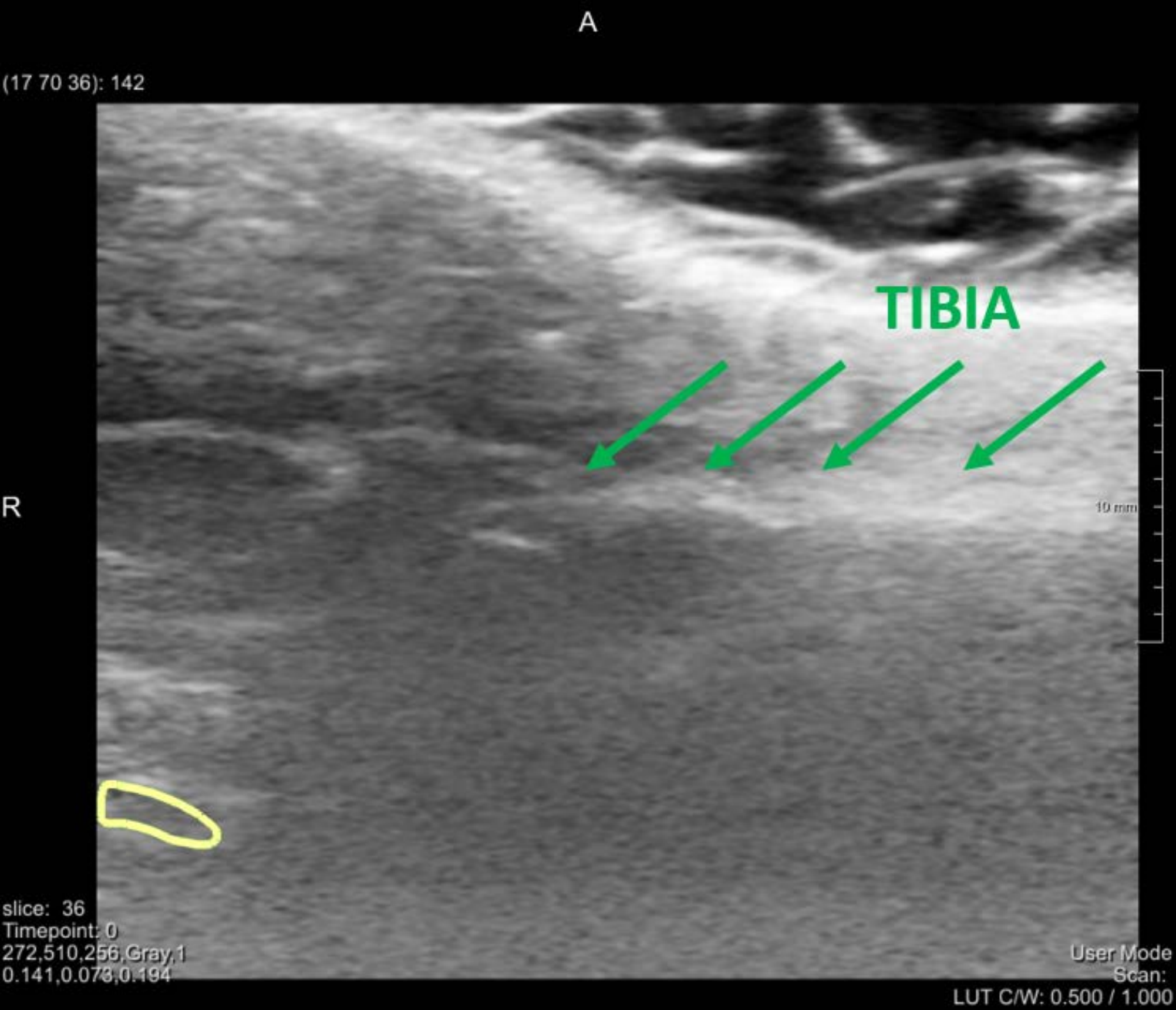
a.

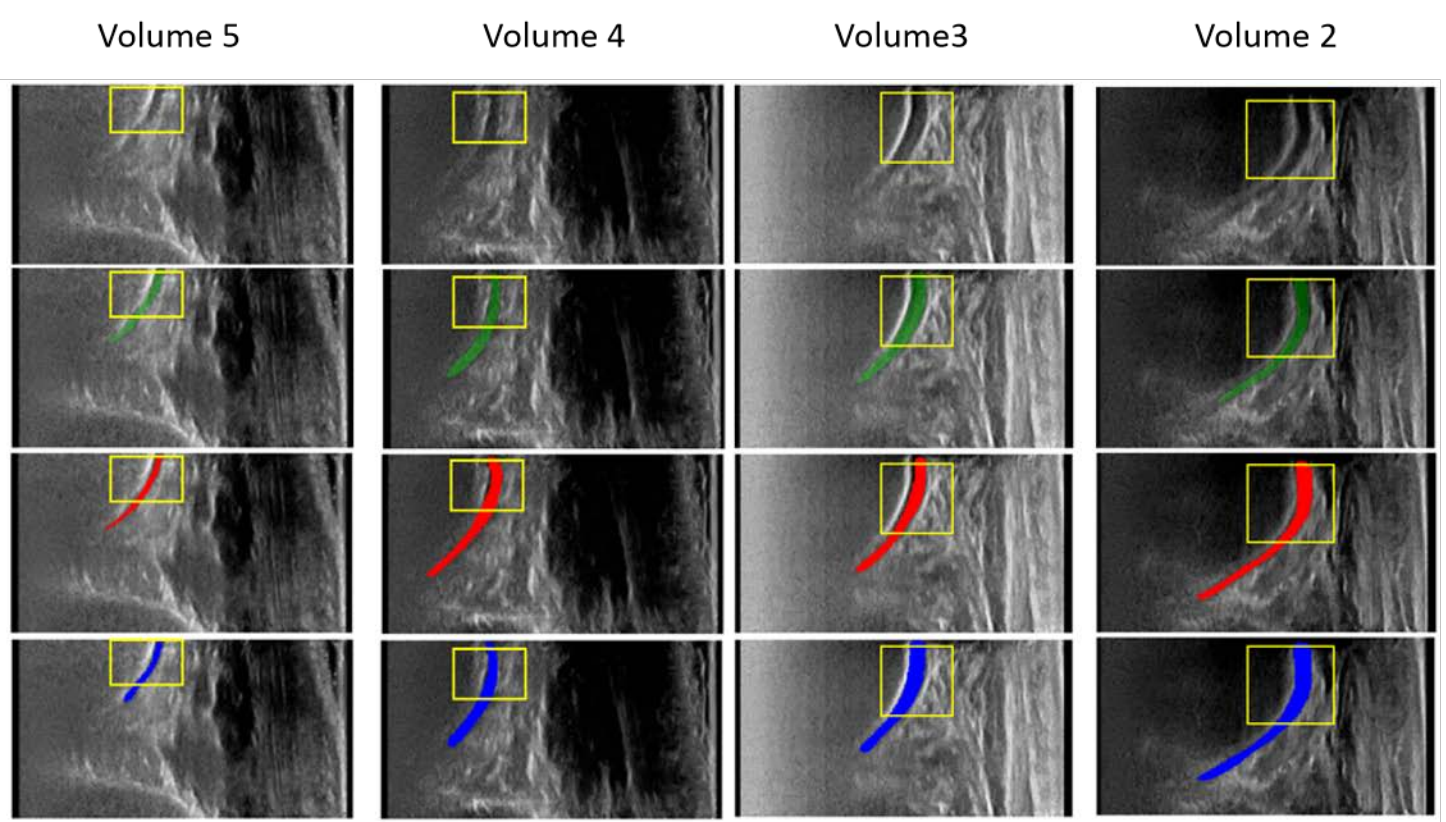


b.


A
(0 154 148): 0
R
20 mm
slice: 148
User Mode
Scan:
LUT C/W: 0.500 / 1.000
a.
A
(2 275 75): 17
ROI
R
20 mm
HYPER-ECHOIC
SUPERIOR AND
INFERIOR BOUNDARIES
slice: 75
User Mode
Scan:
LUT C/W: 0.500 / 1.000
b.




A
(17 70 36): 142
TIBIA
R
10 mm
slice: 36
Timepoint: 0
272,510,256,Gray,1
0.141,0.073,0.194
User Mode
Scan:
LUT C/W: 0.500 / 1.000


Volume 5
Volume 4
Volume3
Volume 2
No contours
UNet
Ground-truth
Intra-observer