Multi-eXpert Fusion: An ensemble learning framework to segment 3d trus prostate images

Clément Beitone, Jocelyne Troccaz

Multi-eXpert Fusion: An Ensemble Learning Framework To Segment 3D TRUS Prostate Images

Clément Beitone* and Jocelyne Troccaz†
Univ. Grenoble Alpes, CNRS, CHU Grenoble Alpes, Grenoble INP, TIMC-IMAG, F-38000 Grenoble, France
(Dated: April 28, 2022)

Purpose: Prostate segmentation of 3D TRUS images is a prerequisite for several diagnostic and therapeutic applications. Unfortunately, this difficult task suffers from high intra- and inter-observer variability, even for experienced urologists/radiologists. This is why automatic segmentation algorithms could have a significant clinical added-value.
Methods: This paper introduces a new deep segmentation architecture consisting of two main phases: view-specific segmentations of 2D slices and their fusion. The segmentation phase is based on three segmentation networks trained in parallel on specific slice viewing directions: axial, coronal, sagittal. The proposed fusion network is then fed with the output of the segmentation networks and trained to produce three confidence maps. These maps correspond to the local trust granted by the fusion network to each view-specific segmentation network. Finally, for a given slice, the segmentation is computed by combining these confidence maps with their corresponding segmentations. The 3D segmentation of the prostate is obtained by re-stacking all the segmented slices to form a volume.
Results: This approach was evaluated on a database of 100 patients with several combinations of network architectures (for both the segmentation phase and the fusion phase) to show the flexibility and reliability of the framework. The proposed approach was also compared to STAPLE, to the majority voting strategy and to a direct 3D approach tested on the same database. The new method outperforms these three approaches on all evaluation criteria. Finally, the results of the Multi-eXpert Fusion (MXF) framework compare favorably with other state-of-the-art methods while these methods typically work on smaller databases.
Conclusions: We proposed a novel MXF framework to segment 3D TRUS images of the prostate. The main feature of this approach is the fusion of expert networks results at the pixel level using computed confidence maps. Experiments conducted on a clinical database have shown the robustness and flexibility of this approach and its superiority over state-of-the-art approaches. Finally, the MXF framework demonstrated its ability to capture and preserve the underlying gland structures, particularly in the base and apex regions.

Keywords: Prostate segmentation, 3D TRUS, Deep Learning, Ensemble learning.

* clement.beitone@univ-grenoble-alpes.fr
† jocelyne.troccaz@univ-grenoble-alpes.fr

I. INTRODUCTION

Prostate cancer is the second most common cancer in men and the fourth most common cancer overall according to the data from GLOBOCAN 2020 [1]. Prostate screening tests may include a Prostate Specific Antigen test (PSA) and a Digital Rectal Examination test (DRE), but only a biopsy examination, most often performed under ultrasound control (Figure I), can certify the presence or absence of cancer cells and allows determining the grade of the cancer.

For this reason, trans-rectal ultrasound (TRUS) examination of the prostate is an essential skill for urologists/radiologists. Accurate segmentation of the prostate from 3D TRUS images serves as the cornerstone of many approaches and therefore has an impact on treatment planning, diagnostic and therapeutic procedures for prostate cancer as well as other prostate diseases. The improvement of this step would, therefore, both benefit the entire treatment chain and contribute to the reliability of the examination as well as to the patient's comfort during it. However, manual segmentation of 3D TRUS images is time-consuming, subjective and with limited reproducibility. It heavily depends on experience and has a large inter- and intra-observer variability [2]. To cope with these issues semi-automatic and automatic algorithms are highly expected by clinical teams.

Several prostate segmentation approaches have been proposed over the past decades [3–5]. Nevertheless most of them deal with the segmentation of 3D MR images where the problem is considered a bit easier. Indeed, on TRUS images, prostate boundaries are often weakly present or absent, especially at the base and apex, and various types of artifacts may be present.

Methods based on deep learning techniques, such as 2D and 3D Convolutional Neural Networks (CNN), have received considerable attention in recent years [6]. As an instance for the 3D approaches Wang *et al.* [7] proposed a 3D CNN based on deep attentive features to segment prostate in 3D TRUS images. Once trained, these approaches are very efficient and provide results almost instantaneously, which again, in the clinical context, translates into shorter time for the examination and therefore less discomfort for the patient. Nevertheless, to be effective, these methods must be trained and evaluated on very large datasets. Moreover, their reliability and potential clinical use is only made possible if these datasets represent the variability of routine clinical cases.

Several 3D US image segmentation approaches use one or more CNN models working on 2D slices of a 3D image and combine the results into a 3D object. One advantage of these approaches is that training CNNs on 2D slices extracted from 3D volumes increases the size of the training database, with each volume generating several dozen of images. Among these 2D approaches, some also combine the result from several views. The rationale is that the visibility of an organ can vary a lot locally: for example, the apex of the prostate is not very visible in axial view but more clearly appears in a sagittal one. These combination techniques can be classified in a framework called ensemble

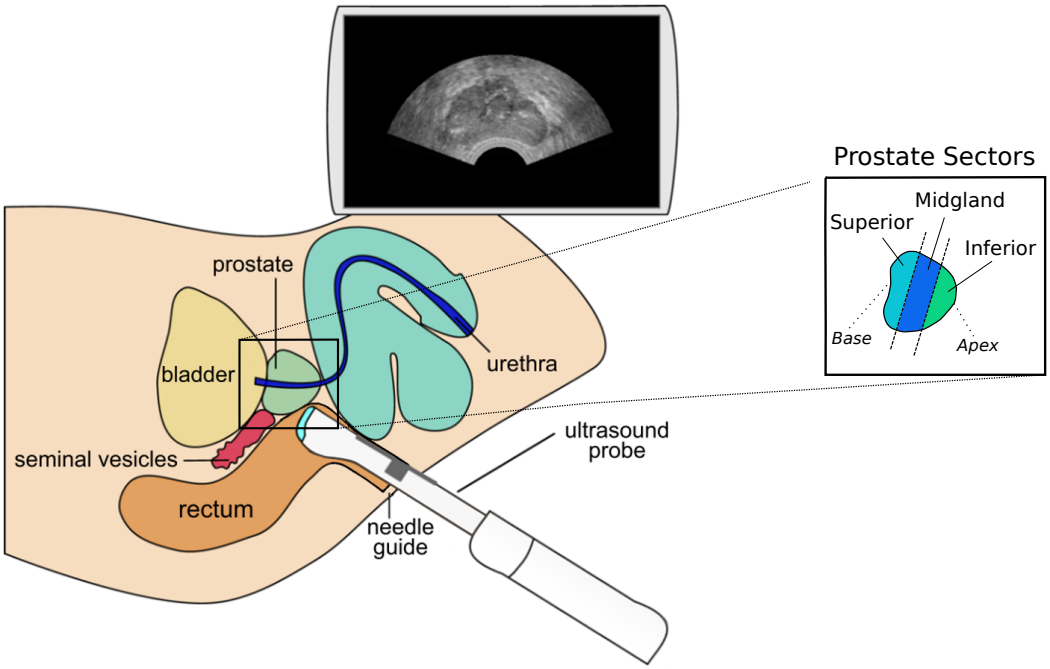

Figure 1. TRUS-guided biopsy procedure - general set-up and typical axial TRUS image

learning [8]. Indeed, it is often found that better performance can be achieved by combining several models in some way, instead of using a single one.

In an ensemble learning strategy, several models, also called experts, are trained in order to partially solve the global problem considered too complex to be solved at once. Then a supervision algorithm is used to combine the *votes* of these models. Such a strategy proves to be reliable and several approaches of this type can be found in the literature; each of them having shown their superiority over direct 3D approaches. For example, Man *et al.* [9] segmented the pancreas with a set of three deformable U-Nets that handles the three main views (axial / coronal / sagittal) of a 3D image and fuses their output with the majority voting strategy. In a majority vote, a voxel is considered to belong to a class if more than fifty percent of the experts agree on this label. With a broader spectrum of application, Perslev *et al.* [10] proposed a generic approach evaluated on ten different segmentation tasks of the Medical Segmentation Decathlon. Their strategy is based on training a U-Net on six different randomly selected views and combining the 6 output volumes using linear fusion. Finally, Orlando *et al.* [11] have developed a deep learning version of a previously proposed work [12] where they merge the segmentation results of a modified U-Net from twelve 2D views from 3D TRUS images and combine them through interpolation to segment the prostate in 3D. Nevertheless, the main drawback of this latter approach lies in the effect of the sampling and smoothing which may have a negative impact on very irregular prostates. More recently, Jiang *et al.* [13] proposed a segmentation average network for the fusion of three 2D segmentation maps in order to segment carotid vessels. Finally, among the fusion approaches, one can also consider the STAPLE (Simultaneous Truth And Performance Level Estimation) approach [14] which allows the fusion of the segmentations given as input in the framework of an expectation maximization algorithm. This approach also allows to assign a score to the experts associated with the inputs. In this framework the probabilistic estimate of the true segmentation is formed by estimating an optimal combination of the segmentations, weighting each segmentation.

The previously described supervision strategy produce results that are then used to globally weight the output of the different models. These weights can be either fixed, as is the case for the majority voting strategy, or variable (*e.g.* linear fusion) when some models are known for their expertise on a subset of the database. Unfortunately, none of these strategies consider that each view-specific network may have a local expertise in a portion of the image. Yet, it is also clear that central slices in each of the three orientations (axial, sagittal, coronal) provide better visualization of the continuity of the prostate contour, supporting the idea that expert networks will be more likely to provide a reliable result on these central slices than on slices located at the ends.

To address this problem, we propose a Multi-eXpert Fusion (MXF) framework that performs a new type of pixel-scale fusion strategy based on CNNs. Similar to a mixture of experts method, a fusion CNN is trained to produce confidence maps, rather than a single scalar weight per expert. These maps correspond to the local confidence granted to each of the view-specific segmentation networks for a given input. We show that this new fusion framework is robust and can be applied on different types of CNN architectures with simiarticlelar outcomes. We discuss quantitatively and qualitatively the results obtained on a large number of patients based on routine clinical data. These results demonstrate the reliability of the proposed approach and allow a fair comparison with the most recent approaches.

II. MATERIALS AND METHODS

A. Dataset and ground truth

The database consists of 100 TRUS volumes from different patients. Data were collected over a three-year period on patients suspected of having a cancer or followed during their treatment. These volumes were acquired with two different motorized 3D end-fire endorectal probes during routine biopsy series using two different commercial guidance systems (Urostation© and Trinity© from Koelis SAS). The examinations were performed by six urologists from the Grenoble University Hospital. No particular constraints or protocols were imposed to the physicians when performing these procedures. These routine acquisitions in a public health institute and over this three-year period ensure the diversity of patients, practitioners, and equipment that make the database representative of the clinical variability of these procedures.

The images were all segmented by the urologists involved in the biopsy procedure using a semi-automatic 3D algorithm (proprietary to Koelis). This segmentation step is composed of three phases: first the expert selects a bounding box around the organ; then the algorithm proposes a segmentation of the organ; finally, the expert can manually modify the computed shape to improve the segmentation. During the second step, the algorithm fits an atlas model of the prostate having been obtained from MRI images [15]. During the third phase, the expert can refine the result produced by the algorithm by adding 3D control points on the ultrasound volume. Finally, before being integrated in the database the 100 US volumes were individually screened by another operator in order to reject the examinations for which the quality of the segmentation was not satisfactory. Data collection was declared to

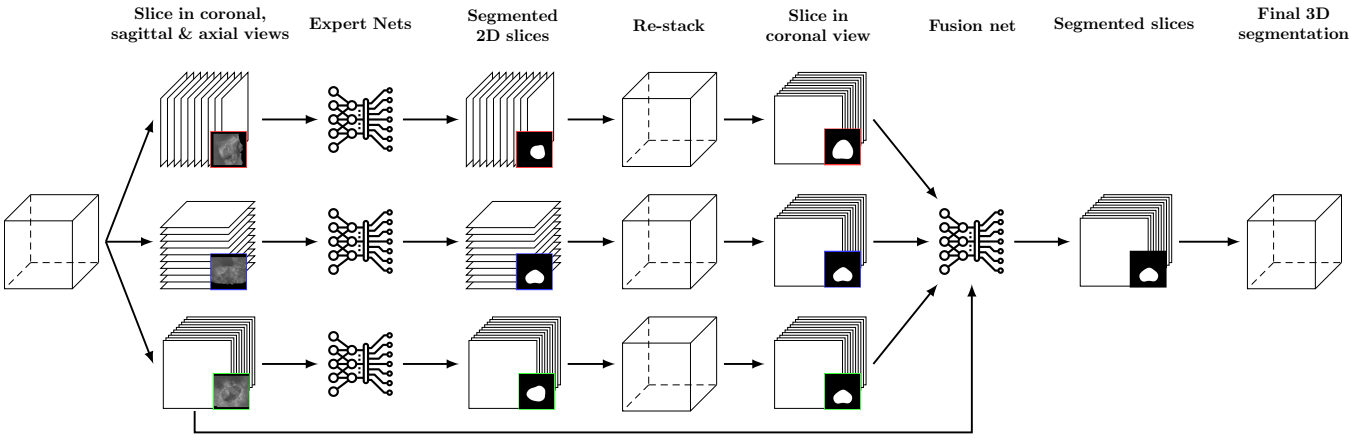

Figure 2. Overall architecture of the MXF framework.

the CNIL (French data protection authority) under the reference MR2711140520. As part of the training procedure, the database was randomly divided into two sets: one for training (80 patients) and one for testing (20 patients). To increase the reliability of the results reported in the following sections, the algorithm was trained with a k-fold validation strategy, with $k = 5$ to keep the 80/20 ratio. All the results provided in the Section III, correspond to the combination of the 20 test subjects from each fold.

B. MXF framework

The overall objective can be stated as finding the prostate surface being given a 3D TRUS input volume X. This requires learning a function F so that the segmented volume $F(X)$ is as close as possible to the ground truth segmentation Y.

Determining this function can be related to a class of methods called *mixture of experts* as described in [8]. In a classical mixture of experts framework, F function can be decomposed into n sub-functions named experts E_i and a supervision function C_i allowing to weight each of the experts' predictions:

$$\hat{Y} = \sum_{i=1}^{n} C_i(X) E_i(X), \text{ with } \begin{cases} 0 \leq C_i(X) \leq 1 \\ \sum_{i=1}^{n} C_i(X) = 1 \end{cases} \quad (1)$$

In the proposed approach, we introduce three experts. Each of them is specialized in the segmentation of 2D US images of a given direction. Each of the three experts (E_i) is a CNN (Section II D) and is trained to segment 2D slices extracted from an input 3D TRUS volume along sagittal, coronal or axial directions (Figure 2). Once the three stacks of 2D slices have been segmented for a given input volume, we can re-stack them to form three segmented volumes.

In a second phase, the initial TRUS volume and these three segmented volumes are fed to a fusion CNN in order to produce the final segmented volume. The originality of the presented approach lies in the fact that the fusion network is trained to produce a tensor of three confidence maps (C_i) used to combine the segmentations from the three experts into a final segmented image ($\hat{Y} = \sum_{i=1}^{3} C_i \times E_i$). Unlike traditional supervision methods, the produced confidence maps allow to take into account the experts' local strengths and weaknesses by assigning them a confidence at the pixel-scale for each input slice. Finally, all the segmented slices are pilled up to form the final segmented volume ($\hat{Y}$) that better estimates the ground truth segmentation.

C. Pre-processing

Prior to the view specific segmentation step, a coarse bounding box is defined around the prostate, using data from the patient's clinical record. This coarse bounding box is the one given by the urologist/radiologist during the routine procedure where data were collected (see Section II A). This information is used to crop the volumes with an additional fixed contextual region of 40 voxels around it. Volumes are then re-sampled to a fixed size (160, 128, 136) which corresponds to an average voxel size of 0.41 mm (±0.05), 0.43 mm (±0.07) and 0.41 mm (±0.06). No other

operation is applied. Since the used training base is already large, containing about 10k 2D slices, we did not observe any improvement using data augmentation techniques (such as adding noise or applying transforms...).

D. Expert networks

Several network architectures have been implemented to perform the 2D expert segmentation phase and thus to assess the reliability and flexibility of the overall proposed framework. The first tested architecture is an encoder/decoder network (EncDec). The second one is a U-Net, which is an EncDec with an additional skip connection between the encoding and decoding blocks. The last one is called residual U-Net and can be considered as a U-Net with additional skip connections within each encoding / decoding block [16].

For the first two architectures (EncDec & U-Net), encoding/decoding blocks are a series of two 3×3 convolutional layers each followed by a batch normalization and a ReLU unit. The encoding blocks double the number of features and are connected by 2×2 max pooling layers. The decoding blocks halve the numbers of features and are connected by a 2×2 up-sampling layers. Finally, a dropout layer has been added after the last encoding block.

The residual U-Net architecture is similar to the U-Net. However, the encoding units are more complex and are composed of two successive sub-units. The first is composed of a series of 1×1 convolution, batch normalization and ReLu. The second is made of two blocks of 3×3 convolution with batch normalization and ReLu. The output of the encoding units is the combination of the output of these two sub-units leading to the short skip connection. Encoding blocks are connected by 2×2 max pooling layers. Decoding blocks include two 3×3 convolutions layers with batch normalization and ReLu. These blocks are connected by transposed convolution layers with batch normalization.

All three architectures take as input a single-channel slice tensor. Their output, which has the same size as the input, is given by a 1×1 convolution followed by a sigmoid unit.

The Adam optimizer was used with a fixed learning rate set at $1e^{-4}$ to optimize a Dice loss function in all scenarios (Equation 2). The Dice loss was chosen since Orlando *et al.* [11] showed that it is an efficient loss for prostate TRUS image segmentation (without significant difference with the Dice / cross-entropy combination).

$$Loss_{Dice}\left(Y, \hat{Y}\right) = 1 - \left(2\frac{|Y \cap \hat{Y}|}{|Y| + |\hat{Y}|}\right) \tag{2}$$

E. Fusion network

Let us remind that the three view-specific experts produce three segmented volumes. To enable the combination of the produced segmentations in the fusion stage, they are re-sliced in the coronal view. The initial volume to be segmented is also re-sliced in the same direction. Thus, the input of the fusion network is a 4-channel tensor composed of a TRUS image in coronal view and three segmented slices obtained at the same location from the three previously view-specific segmented volumes. The network is trained to produce a set of 2D confidence maps C_i as a 3-channel tensor (cf. Figure 3). These three confidence maps reflect the trust that the fusion network assigns locally (pixel-scale) to each of the view-specific segmentations in order to produce the final segmentation.

For the fusion network, we have also tested the network architectures described in Section II D. Apart from the adaptation of the number of input and output channels, the only modification was the replacement of the last layer (sigmoid) by a softmax layer. This is required in order to keep the resulting segmentation (sum of the products cf. Equation 1) in the range $[0, 1]$. Figure 4 illustrates the structure of this fusion network.

As presented in Figure 4, the loss function evaluates the error between the ground truth segmentation Y of the slice and the prediction $\hat{Y}$ computed using the confidence maps and the expert segmentations. Several combinations between two area-specific cost functions (Dice (Equation 2) and Smooth L1 (Equation 3 with $\alpha = 1.0$)) and one cost function based on the Hausdorff distance estimation (HD) have been evaluated for this purpose. The HD loss is based on morphological operations as it was originally proposed by Karimi *et al.* [17] (Equation 4). In Equation 4, $\ominus_k$ denote k successive erosion with the kernel $B = \begin{pmatrix} 0 & 1.5 & 0 \\ 1.5 & 1.5 & 1.5 \\ 0 & 1.5 & 0 \end{pmatrix}$. The Adam optimizer was used with a fixed learning rate set to $1e^{-4}$ to optimize the losses.

$$Loss_{SL1}\left(Y, \hat{Y}\right) = \begin{cases} |Y - \hat{Y}| & \text{if } |Y - \hat{Y}| > \alpha \\ \frac{1}{|\alpha|}(Y - \hat{Y})^2 & \text{if } |Y - \hat{Y}| \leq \alpha \end{cases} \tag{3}$$

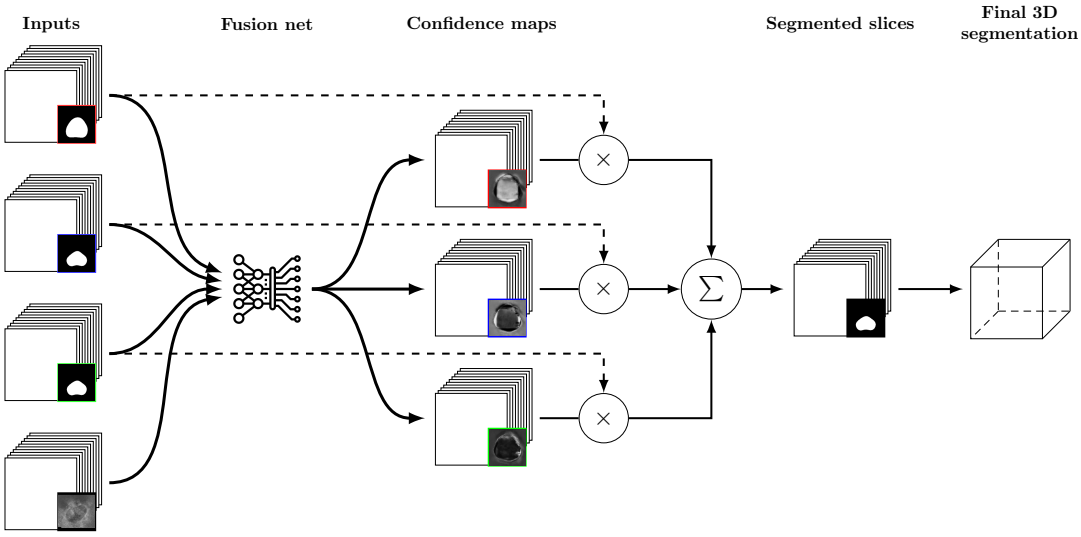

Figure 3. Details of the fusion stage.

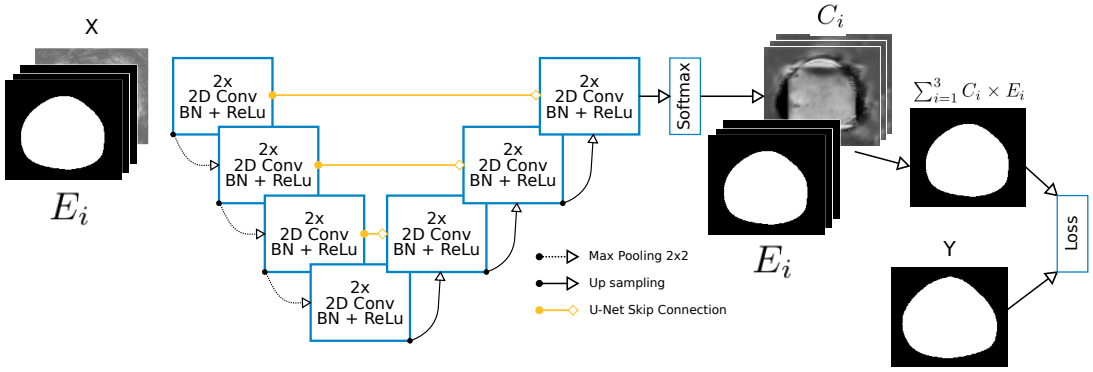

Figure 4. U-Net based version of the fusion network.

$$Loss_{HD}\left(Y, \hat{Y}\right) = \frac{1}{|\Omega|} \sum_{k=1}^{K} \sum_{\Omega} \left(\left(Y - \hat{Y}\right)^2 \ominus_k B \right) k^2 \tag{4}$$

F. Evaluation & implementation

To assess the versatility and reliability of the proposed framework, twelve different scenarios (pair of architecture and loss function) were compared based on the three network architectures and the four loss combinations presented in Sections II D & II E. Within each scenario the same network architecture is used for both the experts and the fusion stage.

In order to give the most comprehensive picture of the quality of the results four 3D metrics are given: two measure the overlapping of segmented surfaces (Dice and Jaccard indexes) and two are related to the distances between contours (average surface distance (ASD) and Hausdorff distance (HD) given in millimeters).

All these networks were implemented within the Pytorch framework and the experiments were conducted on a Titan RTX from NVIDIA. All networks were trained separately: first the three expert networks and then the fusion network. They were trained for 100 epochs with a batch size of 50 2D images.

To demonstrate the impact of the fusion strategy, we also evaluated the performances of each expert individually after having re-stacked the segmented 2D slices (see Figure 2).

In order to compare the proposed approach with state-of-the-art methods, we implemented two baseline fusion

Table I. MXF framework performances for the twelve evaluated scenarios (see Section II F).

Architecture		Encoder/Decoder		Residual U-Net		U-Net	
Criterion \ **Loss**		Dice	Dice+HD	Dice	Dice+HD	Dice	Dice+HD
Dice	mean	0.92	0.92	0.92	0.92	**0.93**	0.92
	std	0.04	0.04	0.04	0.04	0.04	0.04
Jaccard	mean	**0.86**	0.85	0.85	**0.86**	0.85	**0.86**
	std	0.07	0.06	0.07	0.07	0.07	0.07
ASD	mean	0.85	0.86	0.85	0.86	0.84	**0.83**
	std	0.50	0.52	0.47	0.48	0.51	0.44
HD	mean	7.41	7.11	7.41	7.47	6.38	6.15
	std	2.12	2.01	4.87	2.46	3.11	2.87
Criterion \ **Loss**		SL1	SL1+HD	SL1	SL1+HD	SL1	SL1+HD
Dice	mean	0.92	0.92	0.92	0.92	**0.93**	0.92
	std	0.04	0.04	0.04	0.04	0.04	0.04
Jaccard	mean	0.85	0.85	0.85	**0.86**	**0.86**	**0.86**
	std	0.07	0.07	0.06	0.07	0.06	0.07
ASD	mean	0.86	0.84	0.85	0.85	**0.83**	**0.83**
	std	0.51	0.49	0.59	0.53	0.41	0.48
HD	mean	7.36	7.83	6.93	6.79	**5.48**	5.42
	std	2.81	2.91	3.24	3.11	2.66	2.58

approaches: the majority voting strategy (MAJ) and simultaneous truth and performance level estimation (STAPLE). The majority voting strategy was implemented as descibed in [9]: at least two of the three experts must give the same answers to label a pixel as part of the prostate. Regarding STAPLE, the expectation-maximization algorithm that considers a collection of segmentations and computes a probabilistic estimate of the true segmentation, we used its SimpleITK implementation.

We also compared the presented method to a 3D global approach using a V-Net. The V-Net [18] is a fully 3D U-Net, it was implemented and configured as suggested by Isensee *et al.* [19] (data augmentation/loss). This 3D network was implemented in the Pytorch framework and trained on the same training and test datasets as our proposed approach.

III. RESULTS

A. Performances of the MXF framework

The performances of the twelve scenarios (Section II F) are summarized in Table I. These results show that the MXF framework allows to obtain very accurate and consistent results regardless of the scenario considered. It can be seen that the overlap criteria (Dice and Jaccard) are in all cases very similar or even identical on all the scenarios, as well as the average surface distance metric (ASD). These three criteria allow to characterize the regularity and the consistency of the results produced by the method. The Hausdorff measurement informs about the maximum deviations along the surface and once again one can notice that these maximum errors are relatively low, about 5 mm to 7 mm maximum for all the scenarios.

Table II presents the compared performances of each expert to MXF best results.

Finally, although our implementation is focused on method quality assessment and is therefore not optimized to run on a clinical platform, the validation step takes less than a minute to segment all patients in the test database. This computation time includes reading, processing and writing the data. These elements ensure that the proposed approach can run almost instantly on the ultrasound platform once the code is optimized for it.

Table II. MXF framework results compared to expert results alone within the best scenario (U-Net).

Criterion \ Method		MXF	Axial	Coronal	Sagittal
Dice	mean	**0.93**	0.87	0.88	0.89
	std	0.04	0.03	0.04	0.03
Jaccard	mean	**0.86**	0.79	0.81	0.81
	std	0.06	0.04	0.06	0.05
ASD	mean	**0.83**	1.37	1.34	1.39
	std	0.41	0.58	0.52	0.57
HD	mean	**5.48**	10.14	11.12	10.01
	std	2.66	4.12	3.41	3.76

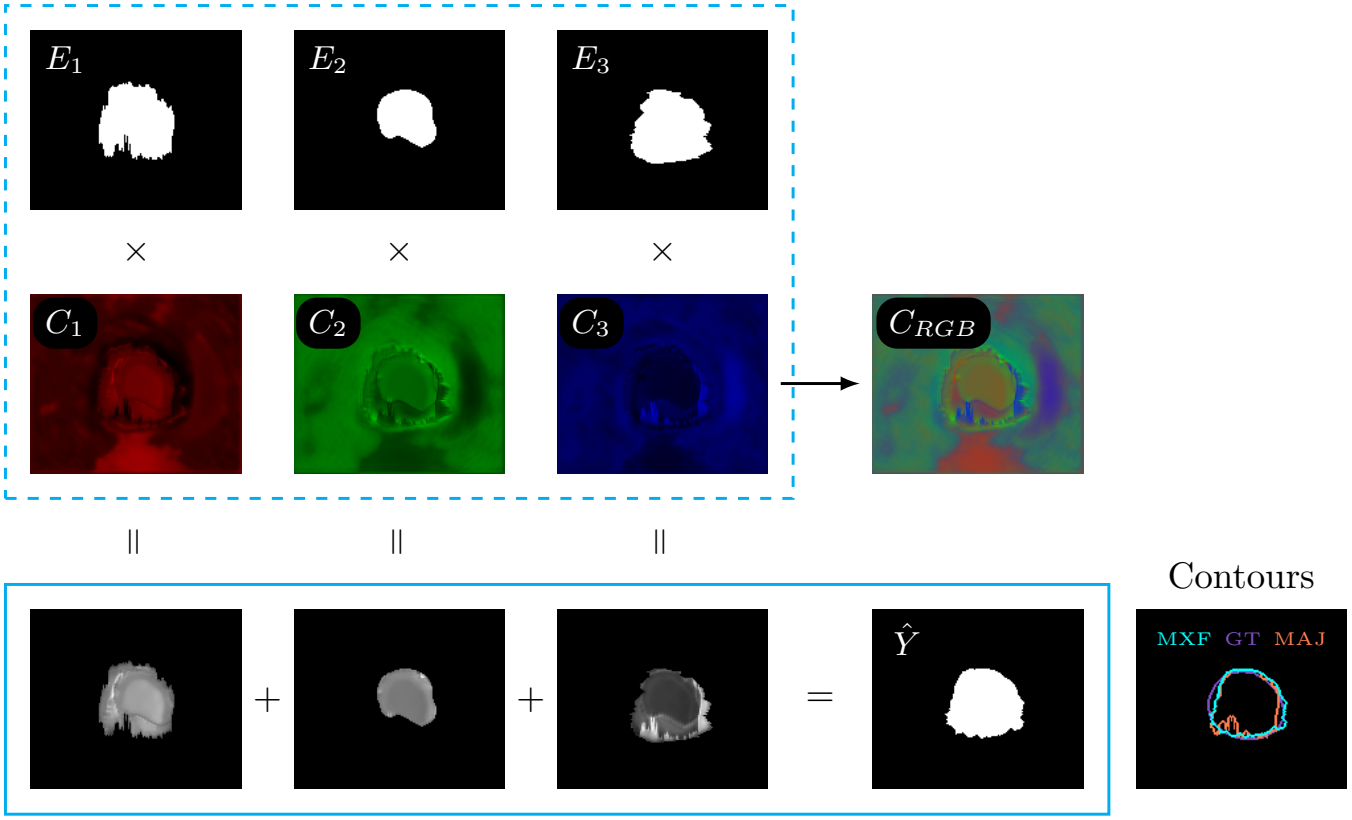

Figure 5. Construction process of the $\hat{Y}$ segmentation from the experts' outputs E_i and the confidence maps C_i provided by the MXF framework, illustrated on a sample slice located close to the apex. Color intensity corresponds to the level of confidence. MXF=our framework; MAJ=majority voting; GT=ground truth

B. Visualization of MXF performances

Figure **??** illustrates the process of building the segmentation of a slice at the end of the MXF pipeline. It shows the type of errors that the majority voting strategy will tend to produce, especially in the inferior region of the prostate (see Figure I) where the experts often disagree as in the presented example. One can see that the proposed method is able to provide a better consensus. The figure also introduces C_{RGB}, a synthetic representation used in Figure 5 to visualize the three confidence maps associated with the three experts as a single RGB image.

Figure 5 gives, for several slices, typical instances of the confidence maps produced by the fusion network and the segmentation obtained from them, for the U-net implementation. These examples show that the contours regularity is much better with the fusion network than with the majority voting strategy, although no regularity constraint was applied. In addition, the fusion network generally produces a single, full and connected element with a rather

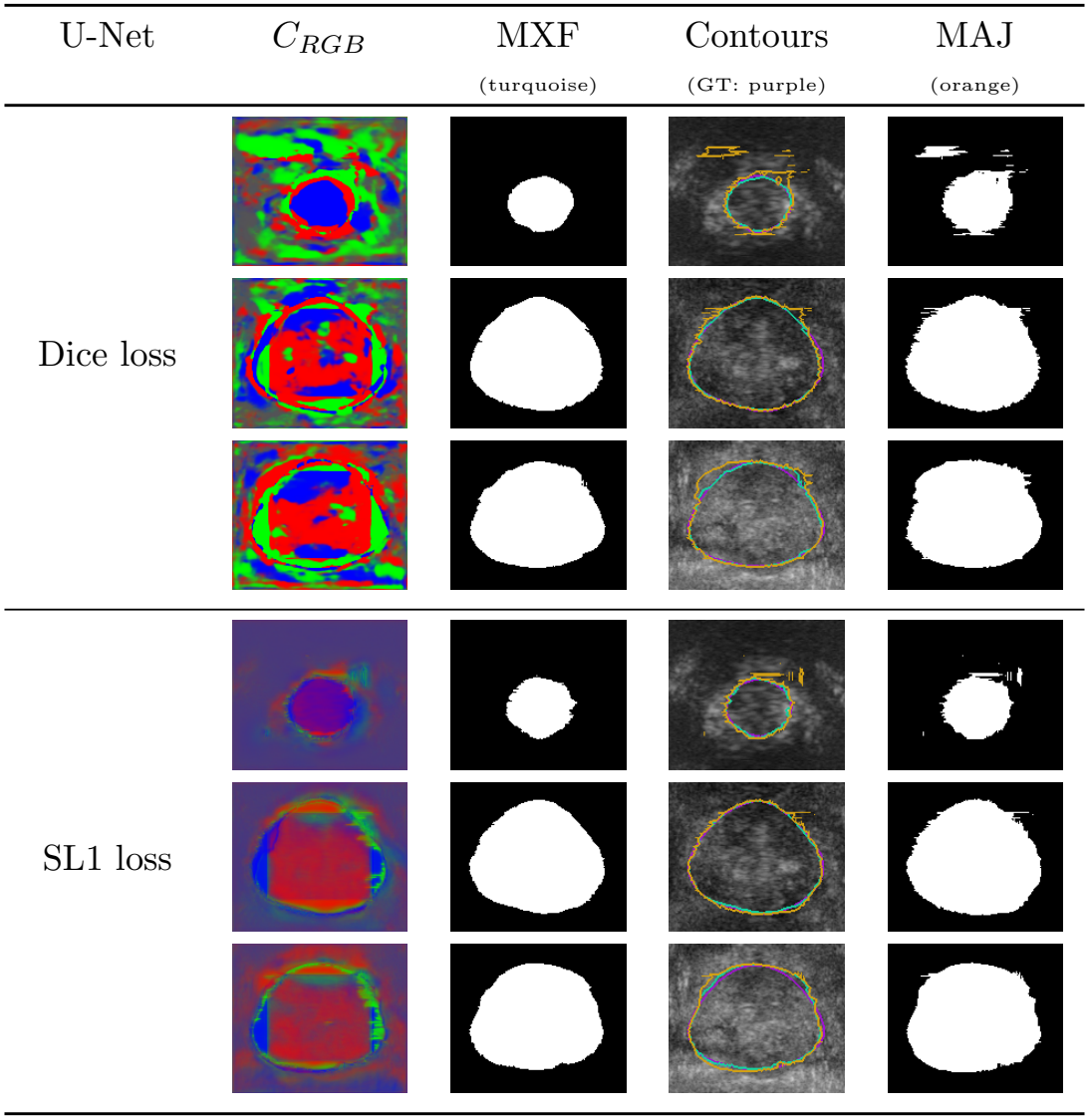

Figure 6. Visual evaluation of MXF results obtained for the U-Net scenario. The three confidence maps C_i are displayed as a single RGB image C_{RGB}, each of the three confidence maps being associated with a color channel (where R, G and B correspond to sagittal, axial and coronal views respectively). The contours on the TRUS image correspond to the ground truth (purple), MXF (turquoise) and majority voting (orange). The figure shows 2D slices from the inferior region (upper row) to the superior region (lower row) of the prostate.

smooth boundary. This figure also shows that, even if the overall results look alike (cf. Table I), the SL1 loss helps to produce smoother confidence maps compared to the Dice loss. This may be related to the fact that this loss function is smoother and explain the results slightly in favor of the SL1 loss (see Section III A).

Figure 6 summarizes the results obtained with the MXF and majority voting fusion strategies (best scenario) on three regions of the gland (from the base to the apex). The regions are computed to represent equivalent sub-volumes of the prostate. It highlights the importance of the fusion network, especially in the prostate regions considered more difficult to segment (base and apex).

C. Comparison with the state of the art

As explained in Section II F, we have implemented the majority voting strategy, the STAPLE strategy and a V-Net to compare our method directly to them. The results obtained with these three approaches are given in Table III.

As it can be seen, the proposed method significantly outperforms the baseline majority voting strategy and STAPLE

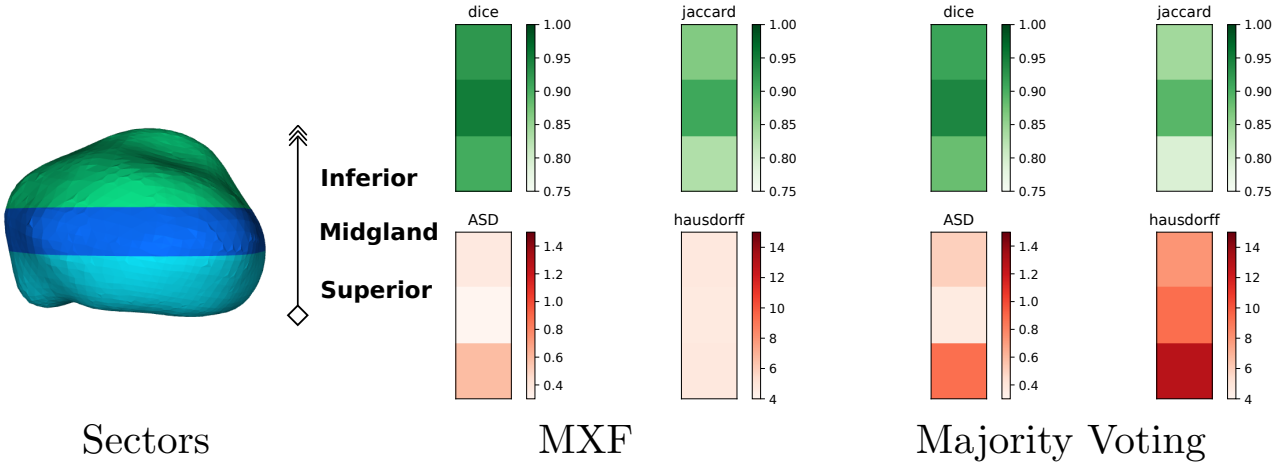

Figure 7. Sectoral evaluation (inferior, midgland, superior as in Figure I) for the MXF framework (best scenario U-Net) and the majority vote strategy. For each sector, for each evaluation criterion, the average value computed on the test dataset is given.

Table III. Comparison of MXF (best scenario U-Net) to a fusion strategy using the majority voting or STAPLE and to a V-Net 3D segmentation (see Section III C).

Criterion \ Method		MXF	MAJ	STAPLE	V-Net
Dice	mean	**0.93**	0.88	0.88	0.89
	std	0.04	0.04	0.04	0.04
Jaccard	mean	**0.86**	0.82	0.82	0.81
	std	0.06	0.06	0.05	0.06
ASD	mean	**0.83**	1.34	1.35	1.41
	std	0.41	0.49	0.49	0.39
HD	mean	**5.48**	6.35	6.46	6.54
	std	2.66	3.02	3.25	2.17

in all scenarios. The results between proposed, majority voting and STAPLE strategies were compared statistically: Shapiro tests within each strategy to demonstrate the normality and Wilcoxon tests to demonstrate significant difference.

Our results can also be favorably compared to the state of the art. For instance, Wang *et al.* [7] obtained an average Dice of 0.90 ± 0.02 (Jaccard 0.82 ± 0.04) and an ASD 1.16 ± 0.7 mm in the best cases on a dataset composed of 40 patients (with a four fold cross validation strategy). Orlando *et al.* [11] have obtained an overall absolute median Dice of 0.94 and an absolute median ASD of 0.89 mm on their test dataset of 20 patients. The corresponding values of our method for the best scenario (U-Net) are 0.93 and 0.84 mm respectively. Finally, Girum *et al.* [20] proposed a 3D segmentation method based on 2D segmentation but without fusion stage and have obtained and average Dice of 0.88 ± 0.02 and an average HD of 8.37 ± 2.93 mm on 14 patients.

IV. DISCUSSION AND CONCLUSION

Our goal is to provide an efficient, reliable and reproducible 3D segmentation method to assist the physician during a clinical examination such as systematic biopsies. The optimal segmentation method should both improve the quality of the medical procedure and limit patient discomfort by shortening intervention time. To ensure the effectiveness of the proposed method, it is necessary to use data that suitably represent the clinical complexity.

We proposed MXF, a novel deep learning-based ensemble learning framework, to segment the prostate gland in 3D TRUS images. The proposed framework consists of two stages. In the first one, three view-specific CNNs, called experts, are trained to segment 2D slices extracted from an input 3D TRUS volume. The second stage is based on a fusion CNN. The three view-specific CNNs are fed by axial, coronal, and sagittal slices, respectively. The fusion

network is fed by axial slices re-sliced from the three segmented volumes reconstructed in the first stage. This fusion CNN is trained to produce three confidence maps used to merge the outputs of these view-specific CNNs at the pixel-scale. Unlike the existing strategies, which reconstruct the volume with fixed weights such as majority voting [9] or without an additional fusion process [11, 20], the presented approach assumes that confidence can be granted at pixel-scale to each of the specific segmentation networks for a given input. This feature is the key contribution of the proposed approach and has not been proposed in an image segmentation framework to the best of our knowledge.

Our database includes 100 patients issued from biopsy examinations routinely performed by the urologists of the University Hospital of Grenoble. The use of data from various machines, different patients and annotated by several clinicians contribute to the robustness of the proposed method. Nevertheless one limitation of the proposed work is the use of a database acquired from a single hospital. In addition, although six experts worked on the segmentation of the US volumes and another operator screened the data to exclude those which did not meet the required quality, each patient volume was segmented by a single urologist. Performing multi-centric evaluation and having multi-operator segmentations for each patient would allow us to better evaluate the inter- and intra-observer variability and increase the robustness of proposed approach.

The evaluation demonstrated that the proposed fusion network approach was able to capture and preserve the underlying structures of the gland, especially in the inferior and superior regions. This ability is very important as these areas, considered to be the most difficult to segment by doctors, are also major parts of the peripheral zone known to be the preferred area for tumor localization. By evaluating scenarios composed of different architecture/loss function pairs, we have also demonstrated the positive contribution and flexibility of our approach and in particular of the fusion stage. Indeed, in each scenario the proposed approach gave better results than the majority voting strategy, STAPLE algorithm and V-net. Although the results obtained with the Dice and SL1 loss functions were not significantly different in quality, it was apparent upon visual inspection that the confidence maps generated by the SL1 loss function appeared smoother.

To put the results obtained by the proposed framework into perspective, the best performances of the view-specific networks (Dice = 0.89, Jaccard = 0.81, ASD = 1.39 mm, HD = 10.01 mm) of the majority voting strategy (0.88, 0.82, 1.34 mm, 6.35 mm) and the STAPLE (0.88, 0.82, 1.35 mm, 6.46 mm), are significantly poorer than the results obtained with the MXF fusion strategy (0.93, 0.86, 0.83 mm, 5.49 mm), especially in terms of ASD and maximum error (Hausdorff criterion). One of the possible reasons behind these results is that the fusion network learns what can be considered as an *a priori*, namely that the slices located in the middle of a slicing axis are easier to segment than the ones at the extremity. The expert networks results reflect these difficulties, that are associated with errors, more or less important, located along the edge of the prostate in different quadrants.

Compared to a 3D approach, like V-Net, the proposed method also performs better, even with the use of data augmentation and specific configurations to set up the 3D CNN and its training environment. The results obtained from our database, generally larger than those found in the literature, can be compared favorably with the other methods from the literature.

From the point of view of calculation time, even if the training of the different networks takes few hours, the application to a new patient is almost instantaneous. This, together with the good results obtained by the method, suggest a possible transition to routine clinical application. Future work will focus on evaluating an even larger database from multiple hospitals and will also seek to assess the impact of inter- and intra-observer variability in expert segmentation on training quality and generalization quality. Finally, the extension of MXF to the segmentation of other non-isotropic modalities (such as MRI) could be an interesting addition to this work.

ACKNOWLEDGMENTS

This work was partly supported by Auvergne Rhône-Alpes region (project ProNavIA) and by the French "Agence Nationale de la Recherche", Investissement d'Avenir program (grants MIAI@Grenoble-Alpes under reference ANR-19-P3IA-0003 and CAMI Labex under reference ANR-11-LABX-0004). The authors also want to thank the urologists from Grenoble University Hospital for providing the data.

The authors have no conflicts to disclose.

[1] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries, CA: A Cancer Journal for Clinicians **71**, 209–249 (2021).

[2] S. Tong, H. N. Cardinal, R. F. McLoughlin, D. B. Downey, and A. Fenster, Intra-and inter-observer variability and

reliability of prostate volume measurement via two-dimensional and three-dimensional ultrasound imaging, Ultrasound in medicine & biology **24**, 673–681 (1998).

[3] X. Yang and B. Fei, 3D prostate segmentation of ultrasound images combining longitudinal image registration and machine learning, Medical Imaging 2012: Image-Guided Procedures, Robotic Interventions, and Modeling **8316**, 83162O (2012).

[4] X. Yang, P. J. Rossi, A. B. Jani, H. Mao, W. J. Curran, and T. Liu, 3D transrectal ultrasound (TRUS) prostate segmentation based on optimal feature learning framework, in *Medical Imaging 2016: Image Processing*, 2016.

[5] D. Karimi, Q. Zeng, P. Mathur, A. Avinash, S. Mahdavi, I. Spadinger, P. Abolmaesumi, and S. Salcudean, *Accurate and Robust Segmentation of the Clinical Target Volume for Prostate Brachytherapy*, volume 11073 LNCS, Springer International Publishing, 2018.

[6] G. Haskins, U. Kruger, and P. Yan, Deep learning in medical image registration: a survey, Machine Vision and Applications (2020).

[7] Y. Wang, D. Ni, H. Dou, X. Hu, L. Zhu, X. Yang, M. Xu, J. Qin, P. A. Heng, and T. Wang, Deep Attentive Features for Prostate Segmentation in 3D Transrectal Ultrasound, IEEE Transactions on Medical Imaging **38**, 2768–2778 (2019).

[8] M. Svensén, C. M. Bishop, Bharadwaj, K. B. Prakash, and G. R. Kanagachidambaresan, *Pattern recognition and machine learning*, Springer, 2007.

[9] Y. Man, Y. Huang, J. Feng, X. Li, and F. Wu, Deep Q Learning Driven CT Pancreas Segmentation With Geometry-Aware U-Net, IEEE Transactions on Medical Imaging **38**, 1971–1980 (2019).

[10] M. Perslev, E. B. Dam, A. Pai, and C. Igel, One Network to Segment Them All: A General, Lightweight System for Accurate 3D Medical Image Segmentation, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) **11765 LNCS**, 30–38 (2019).

[11] N. Orlando, D. J. Gillies, I. Gyacskov, C. Romagnoli, D. D'Souza, and A. Fenster, Automatic prostate segmentation using deep learning on clinically diverse 3D transrectal ultrasound images, Medical Physics **47**, 2413–2426 (2020).

[12] W. Qiu, J. Yuan, E. Ukwatta, and A. Fenster, Rotationally resliced 3D prostate TRUS segmentation using convex optimization with shape priors, Medical Physics (2015).

[13] M. Jiang, J. D. Spence, and B. Chiu, Segmentation of 3D ultrasound carotid vessel wall using U-Net and segmentation average network, 42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC) , 2043–2046 (2020).

[14] S. K. Warfield, K. H. Zou, and W. M. Wells, Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation, IEEE transactions on medical imaging **23**, 903–921 (2004).

[15] S. Martin, J. Troccaz, and V. Daanen, Automated segmentation of the prostate in 3D MR images using a probabilistic atlas and a spatially constrained deformable model, Medical physics **37**, 1579–1590 (2010).

[16] L. Yu, X. Yang, H. Chen, J. Qin, and P. A. Heng, Volumetric convnets with mixed residual connections for automated prostate segmentation from 3d MR images, 31st AAAI Conference on Artificial Intelligence, AAAI 2017 , 66–72 (2017).

[17] D. Karimi and S. E. Salcudean, Reducing the Hausdorff Distance in Medical Image Segmentation with Convolutional Neural Networks, IEEE Transactions on Medical Imaging **39**, 499–513 (2020).

[18] F. Milletari, N. Navab, and S. A. Ahmadi, V-Net: Fully convolutional neural networks for volumetric medical image segmentation, Proceedings - 2016 4th International Conference on 3D Vision, 3DV 2016 , 565–571 (2016).

[19] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation, Nature Methods **18**, 203–211 (2021).

[20] K. B. Girum, A. Lalande, R. Hussain, and G. Créhange, A deep learning method for real-time intraoperative US image segmentation in prostate brachytherapy, International Journal of Computer Assisted Radiology and Surgery (2020).