Medical Image Analysis 57 (2019) 186–196

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Accurate and robust deep learning-based segmentation of the prostate clinical target volume in ultrasound images

Davood Karimi[a,*], Qi Zeng[a], Prateek Mathur[a], Apeksha Avinash[a], Sara Mahdavi[b], Ingrid Spadinger[b], Purang Abolmaesumi[a], Septimiu E. Salcudean[a]

[a] Department of Electrical and Computer Engineering, University of British Columbia, Vancouver, BC, Canada
[b] Vancouver Cancer Centre, Vancouver, BC, Canada

ARTICLE INFO

Article history:
Received 21 February 2019
Revised 6 June 2019
Accepted 4 July 2019
Available online 15 July 2019

MSC:
41A05
41A10
65D05
65D17

Keywords:
Image segmentation
Model uncertainty
Shape models
Clustering
Deep learning

ABSTRACT

The goal of this work was to develop a method for accurate and robust automatic segmentation of the prostate clinical target volume in transrectal ultrasound (TRUS) images for brachytherapy. These images can be difficult to segment because of weak or insufficient landmarks or strong artifacts. We devise a method, based on convolutional neural networks (CNNs), that produces accurate segmentations on easy and difficult images alike. We propose two strategies to achieve improved segmentation accuracy on difficult images. First, for CNN training we adopt an adaptive sampling strategy, whereby the training process is encouraged to pay more attention to images that are difficult to segment. Secondly, we train a CNN ensemble and use the disagreement among this ensemble to identify uncertain segmentations and to estimate a segmentation uncertainty map. We improve uncertain segmentations by utilizing the prior shape information in the form of a statistical shape model. Our method achieves Hausdorff distance of 2.7 ± 2.3 mm and Dice score of $93.9 \pm 3.5\%$. Comparisons with several competing methods show that our method achieves significantly better results and reduces the likelihood of committing large segmentation errors. Furthermore, our experiments show that our approach to estimating segmentation uncertainty is better than or on par with recent methods for estimation of prediction uncertainty in deep learning models. Our study demonstrates that estimation of model uncertainty and use of prior shape information can significantly improve the performance of CNN-based medical image segmentation methods, especially on difficult images.

1. Introduction

1.1. Motivation and background

Prostate cancer is the second most diagnosed and sixth deadliest cancer among men world-wide (Jemal et al., 2011). There exist a wide range of prostate cancer management and treatment options including active surveillance, radiation therapy, chemotherapy, hormone therapy, and radical prostatectomy. Prostate brachytherapy is a form of radiation therapy, wherein a number of small radioactive seeds are implanted inside the prostate gland, with the goal of treating the entire prostate gland, regardless of where the tumor is. Due to its targeted nature, brachytherapy is expected to reduce the exposure of the healthy tissue to harmful radiation. Therefore, it is known as an effective treatment option for localized prostate cancer (Morris et al., 2013).

Prior to the placement of brachytherapy seeds, the shape and size of the prostate gland is determined using trans-rectal ultrasound (TRUS) imaging. The acquired images usually consist of a series of 7 to 14 2D ultrasound images that cover the entire prostate gland from the base to the apex. A clinical target volume (CTV), which closely follows the prostate boundary, is delineated on these images (Salembier et al., 2007). This CTV is slightly dilated by an expert radiation oncologist to form a planning target volume (PTV), which is used to decide on the desired distribution of the brachytherapy seeds (Sylvester et al., 2009). Accurate delineation of CTV is critically important because it will help ensure maximum radiation is delivered to the desired locations while minimizing the radiation exposure of the healthy tissue and the surrounding anatomy.

Accurate segmentation of the CTV is a very challenging task because image landmarks are often weak or non-existent, especially at the prostate base and apex. Moreover, strong speckle noise and

* Corresponding author.
E-mail address: karimi@ece.ubc.ca (D. Karimi).

https://doi.org/10.1016/j.media.2019.07.005
1361-8415/

various types of artifacts such as micro-calcifications, ultrasound shadowing, and reverberation can be present. Therefore, manual segmentation is tedious and prone to high inter-observer variability.

Automatic segmentation methods can potentially improve the speed and reproducibility of segmentation. Furthermore, if they are fast enough to perform the segmentation in real time, they can be used to segment the CTV during the seed implantation. This can help account for the changes in prostate shape caused by patient positioning, bladder and rectum filling, and deformations due to the force exerted by the TRUS probe. Such intra-operative adjustments in the seed placement will allow for delivery of the prescribed radiation dose to the CTV (Nag et al., 2000).

1.2. Related works

Many semi-automatic and fully-automatic algorithms have been proposed for segmentation of the prostate or the CTV in ultrasound images. In general, compared with other imaging modalities such as magnetic resonance imaging (MRI) and computed tomography (CT), there has been a greater reliance on prior knowledge in the form of shape and appearance models for segmentation of prostate in ultrasound images (Ghose et al., 2012; Mozaffari and Lee, 2016; Noble and Boukerroui, 2006). This is because the intrinsic image features in ultrasound images are often insufficient for accurate localization and delineation of the prostate boundary. Several studies have proposed methods based on deformable shape models for prostate segmentation in TRUS. These include level-set methods (Li et al., 2016; Kachouie et al., 2006) as well as methods based on active contour models (Jendoubi et al., 2004; Zaim and Jankun, 2007) and active shape models (Betrouni et al., 2005; Hodge et al., 2006). One study suggested adaptively refining the shape model during segmentation for each patient in order to take into account the inter-patient variability in the prostate shape (Yan et al., 2011). These methods are driven by energy functions that depend primarily on the object edge information, which can be problematic for prostate segmentation in TRUS due to the lack of strong edges and presence of artifacts. Therefore, these methods often require elaborate image pre-processing steps to detect and amplify the prostate boundary. Furthermore, most of these methods also rely on a good initialization, and hence need input from a human expert.

A number of studies have exploited the relatively low variability in the prostate shape to suggest segmentation methods that fit a particular parametric shape to the image intensity data. Ellipses, ellipsoids, and super-ellipsoids are the most commonly used shapes (Saroul et al., 2008; Mahdavi et al., 2011). These methods typically require that several control points be manually specified to initialize or anchor the shape to be fitted. Therefore, these methods are not fully automatic. Moreover, to achieve high segmentation accuracy, many of these methods rely on post-processing of the segmentation obtained with shape fitting.

Another class of methods includes those based on machine learning. Methods that rely on shape statistics learned from the training data can be considered as machine learning methods (Qiu et al., 2015; Yang and Fei, 2012). Some studies combine statistical shape models with various machine learning techniques such as random forests, probabilistic models, and dictionary learning, for segmentation (Ghose et al., 2013; Nouranian et al., 2015; 2016). More recently, deep learning methods and in particular convolutional neural networks (CNNs) have been successfully applied for segmentation of prostate or CTV in TRUS images (Anas et al., 2017; Wang et al., 2018; Zeng et al., 2018; Gibson et al., 2018; Ghavami et al., 2018).

From the above review, most of the existing methods have one of several shortcomings. Many of them require careful initialization, usually by a human expert. Also, most of these methods are too slow for real-time segmentation. Moreover, although some of the published methods have reported good average segmentation performance in terms of such criteria as the Dice Similarity Coefficient (DSC), worst-case performance measures such as the Hausdorff Distance (HD) are either not reported or display large variances. This is because, as shown in the examples in Fig. 1, some TRUS images can be particularly difficult to segment due to weak prostate edges and strong artifacts. Such images also pose a challenge for deep learning-based methods that are among the best methods for medical image segmentation and the focus of the present study. Because deep learning models have a high representational capacity and are trained using stochastic gradient descent with a uniform sampling of the training data, their training is easily dominated by the more typical samples in the training set, leading to poor generalization on more challenging but less-represented cases.

1.3. Prediction uncertainty in deep learning models

One of the unique aspects of our methods proposed in this paper compared with previous studies on automatic CTV or prostate segmentation in TRUS is our attention to segmentation uncertainty. This is an important consideration because, as we mentioned above, the range of useful landmarks varies greatly among different TRUS images. Even within the same image, different parts of the image may be quite different in terms of the presence of genuine landmarks and artifacts or noise. Hence, understanding where the segmentations are more reliable can be quite useful. In this section, we review the recent studies on estimating the prediction uncertainty in deep learning models for computer vision.

Certainty has been a long-standing concern in computer vision (Blake et al., 1993; Barra and Boire, 2001). In this regard, deep learning-based methods, which have achieved record-breaking performance in many computer vision applications, have received particular attention (Kendall and Gal, 2017; Kendall and Cipolla, 2016; Lakshminarayanan et al., 2017; Pawlowski et al., 2017). Studies have shown that, unlike older neural network models with a small number of layers, deep neural networks with tens of layers are poorly calibrated (Guo et al., 2017). Here, a classifier is said to have a calibrated confidence if the probability that it assigns to the predicted class reflects its likelihood of being correct. For a perfectly-calibrated classifier, $P\left(y_{\text{predicted}} = y_{\text{true}} | \hat{p} = p\right) = p$, where $\hat{p}$ is the probability of predicting the correct class. It has been shown that deep learning models produce highly over-confident predictions (Guo et al., 2017).

This is a very important topic as deep learning models are employed in a growing range of applications. Predicting the wrong class with high confidence has already led to disastrous consequences in some real-world applications of deep learning models (Kendall and Gal, 2017; Guo et al., 2017). If the model is well-calibrated, on the other hand, the likelihood of such disasters can be reduced by identifying uncertain predictions and asking another model or a human to intervene. In medical applications where the health of patients is at stake, an estimate of the reliability of the predictions generated by a computerized method can be extremely useful. Moreover, modeling the prediction uncertainty may also lead to more accurate models (Kendall and Gal, 2017).

Empirically, it has been shown that increasing the model size (depth and width) and batch normalization worsen the miscalibration of deep neural networks, whereas weight decay alleviates it Guo et al. (2017). It has also been shown that the choice of the loss function used for training a deep learning model has an impact on the calibration of its predictions (Guo et al., 2017; Lakshminarayanan et al., 2017). However, a deeper understanding of the nature of this problem is more difficult. Early efforts to understand uncertainties of neural networks were based on Bayesian

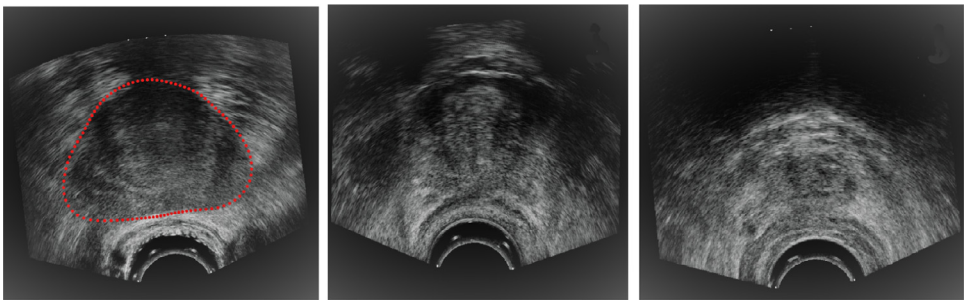

Fig. 1. The left image is an example of a 2D TRUS image with strong prostate boundaries. The other two images show examples with weak or incomplete edge information.

methods (Neal, 2012; Gal, 2016). Although Bayesian models are easy to formulate, they require substantial modifications to standard neural network models and training procedures and can be computationally prohibitive. Consequently, recent studies have developed approximate Bayesian methods or non-Bayesian approaches to estimate model uncertainties. A good example of these methods is the dropout variational inference (Gal and Ghahramani, 2015). This method is based on training the model with dropout regularization (Srivastava et al., 2014) and then applying dropout at test time to sample from the approximate posterior. Building on this method, another study has suggested effective methods for estimating both epistemic and aleatoric uncertainties (Kendall and Gal, 2017). Epistemic uncertainty, also known as model uncertainty, is the uncertainty in the model, e.g., the uncertainty in the weights of a neural network. This certainty can be modeled by considering a distribution on the network weights. Aleatoric uncertainty, on the other hand, is the uncertainty caused by the noise in the observations. It is modeled via probability distributions on the model outputs. Whereas epistemic uncertainty is reduced with more training data, aleatoric uncertainty is a function of each individual data sample. In computer vision applications with huge amounts of labeled training data, it has been shown that modeling aleatoric uncertainty is very useful, but modeling aleatoric uncertainty alone leads to models that are unable to identify novel data that are different from the training set (Kendall and Gal, 2017).

To estimate the epistemic and aleatoric uncertainties for a classification problem, it has been proposed to consider a probability distribution over the logits (Kendall and Gal, 2017). Specifically, for an input x_i, it is assumed that the logits are modeled as:

$$\hat{y}_i|W \sim \mathcal{N}\left(f_i^W, \left(\sigma_i^W\right)^2\right) \tag{1}$$

Where W denote the network weights. The class probabilities are then obtained as $\hat{p}_i = \text{Softmax}(\hat{y}_i)$. Both f and σ are predicted by the same neural network. What is interesting about this model is that ground-truth uncertainty estimates are not needed for the training data. An approximation of the expected log-likelihood is obtained via Monte Carlo integration and sampling the logits' distribution:

$$\begin{aligned} L_W &= \sum_i \log \frac{1}{T} \sum_{t=1}^{T} \exp\left(\hat{y}_{i,t,C} - \log \sum_c \exp \hat{y}_{i,t,C}\right) \\ \text{where:} \quad & \hat{y}_{i,t} = f_i^W + \sigma_i^W u_t \qquad u_t \sim \mathcal{N}(0, I) \end{aligned} \tag{2}$$

In the above equation, T denotes the number of dropout masks simulated. Then for an input x, the prediction uncertainty resulting from epistemic uncertainty in the model weights is obtained through Monte Carlo dropout:

$$p(y = c|x, X, Y) \approx \frac{1}{T} \sum_{t=1}^{T} \text{Softmax}\left(f^{\hat{W}_t}(x)\right) \tag{3}$$

where $\hat{W}_t$ is the model weight matrices sampled from the dropout distribution. The prediction uncertainty can then be estimated as the entropy of p, i.e., $H(p) = \sum_c p_c \log(p_c)$ Kendall and Gal (2017).

Another suggested method for estimating the prediction uncertainty of deep learning models is a post-processing method based on Platt scaling (Guo et al., 2017). In this method, after the main model training is completed, a simple model, $p_i = \sigma(af_i + b)$, is trained to map the model logits to probabilities. Here, a and b are scaling parameters and σ is the sigmoid function. In order to obtain a well-calibrated model, a and b are learned by optimizing the negative log-likelihood on a separate validation set. Furthermore, a recent study proposed a methodology for estimating prediction uncertainty that consisted of using a proper scoring rule, training on adversarial examples, and using an ensemble of models (Lakshminarayanan et al., 2017). The idea of using a model ensemble in that paper is similar to our approach. However, that work used the negative log-likelihood (NLL) for uncertainty estimation and a simple random shuffling of the training data. By contrast, we propose a completely different formulation based on the Kohavi-Wolpert variance (Kohavi et al., 1996) for estimating uncertainty and a more sophisticated and novel approach for sampling the training data.

In this work, we are interested in estimating the uncertainty in the segmentation of the CTV in TRUS images. For this, we propose our own method, which is based on the disagreement among an ensemble of CNNs trained on different subsets of the training data. We explain our method in Section 2. For a comparison with the existing methods, we compare our method with the methods proposed by Kendall and Gal (2017) and Guo et al. (2017), which we will refer to as Epistemic-Aleatoric method and Platt scaling method, respectively.

1.4. Contributions of this study

Existing methods for segmentation of the prostate/CTV in TRUS images, reviewed in Section 1.2, do not offer a way of identifying images that are difficult to segment and they treat all images in the same way. In this paper, we propose a method for segmentation of the CTV in 2D TRUS images that aims at achieving accurate segmentation on easy as well as difficult images while at the same time reducing large segmentation errors. We achieve this by a combination of techniques to identify and pay more attention to difficult images during training, identify difficult images in the test set by estimating the segmentation uncertainty, and by using the prior shape information to improve the uncertain segmentations. There are three main novel aspects to the segmentation method proposed in this work.

1. We propose a new CNN architecture for prostate CTV segmentation in TRUS images. Our proposed architecture computes multi-scale features directly from the input image. These features are merged to predict the CTV segmentation. This architecture improves the segmentation accuracy compared with standard architectures.
2. We propose a method for adaptive sampling of the training images in order to drive the trained model towards achieving more accurate segmentations on difficult images. To this

end, we first learn similarities among all training images based on features learned by a Convolutional Auto-Encoder (CAE). Then, during CNN training with a cross-validation approach, we preferentially sample those training images that are more similar to difficult images in a validation set.
3. We estimate the uncertainty of segmentation predictions and use the prior knowledge regarding the expected shape of the CTV to improve uncertain segmentations. In order to estimate the segmentation uncertainty, we train an ensemble of segmentation models and propose to estimate a segmentation uncertainty map based on the degree of disagreement among these models. We propose a novel method to improve uncertain segmentations based on the estimated uncertainty map and the expected shape.

The initial results of this work were presented as a conference paper (Karimi et al., 2018). The current manuscript extends that conference paper in several ways.

1. We include more review of the literature to place our work within the context of existing methods. We present our methods in detail and a set of comprehensive results that could not be included in our conference paper because of the page limit.
2. We review recent studies on estimating the prediction uncertainty in deep neural networks. We compare our proposed method for estimating the segmentation uncertainty with two recent methods mentioned above.
3. We present a comprehensive comparison of our methods with existing techniques. In the initial paper we compared our CNN architecture with U-net, which was a general-purpose architecture. Here, we also compare with a more recent architecture based on deep attentional features that has been proposed specifically for prostate segmentation in ultrasound images (Wang et al., 2018). Furthermore, we compare our proposed SSM-based method for improving uncertain segmentations with Simultaneous Truth and Performance Level Estimation (STAPLE) (Warfield et al., 2004).

2. Materials and methods

2.1. Data

The data used in this work consisted of the B-mode axial TRUS images of 675 brachytherapy patients. These images were acquired prior to brachytherapy. From each patient, 7 to 14 2D TRUS images were acquired. Each image was 415×490 pixels in size, with a pixel size of $0.15 \times 0.15\text{mm}^2$. All images were collected using BK Pro-Focus or BK Flex Focus machines with a BK biplane 8848 probe (BK Medical, Herlev, Denmark).

A side-firing transrectal probe was used to collect the images at axial intervals of 5 mm, covering the prostate gland from the base to the apex. A semi-automatic method was used to segment the CTV in each volume. This method has been described in detail in Mahdavi et al. (2011). It takes advantage of both the expert knowledge and prior shape information. The method is based on fitting a tapered warped ellipsoid to the image. The method is initialized by a human expert defining the locations of six key-points on a mid-gland slice image. The segmentation pipeline involves un-warping of the image to compensate for the effect of the TRUS probe, mid-gland ellipse fitting, propagation of the mid-gland slice contour to other slices, and finally ellipsoid fitting and image warping. This method is used routinely in the brachytherapy workflow at the Vancouver Cancer Centre. The segmentations produced by this semi-automatic software are manually adjusted by an experienced radiation oncologist to obtain the CTV, which we use as the gold standard in this study. Experiments by Mahdavi et al. (2011) have shown that the ranges of intra- and inter-observer variability in whole-gland non-overlapping volume error for this task are approximately $4-6\%$. When using the semi-automatic software followed by expert modifications, these variabilities are reduced to approximately $3-3.5\%$. Furthermore, they showed that the segmentation errors by this method are within the range of manual intra-observer and inter-observer variabilities.

2.2. The proposed segmentation method

This section explains our proposed segmentation method. We first describe our approach to identifying similar images in the training set and our proposed CNN architecture. Then, we outline our training strategy and explain our methods to detect and improve highly uncertain segmentations.

2.2.1. Clustering of the training images

As we will explain below, our training strategy aims at achieving high accuracy on difficult images. To this end, we identify the difficult images in a validation set and samples the training images based on how likely they are to contribute to improving the segmentation of those difficult images. Therefore, our proposed framework will require a method to quantify image similarities. In this work, we use the sparse subspace clustering (Elhamifar and Vidal, 2013) for this purpose. As suggested by Ji et al. (2017), we apply this method on features learned with a CAE. A schematic depiction of the CAE architecture is shown in Fig. 2. The low-dimensional image representation learned by the CAE is denoted with z^i_{enc}. This representation is mapped into the input to the decoder, denoted with z^i_{dec}, using a fully-connected layer is represented with a matrix, Γ. To have a linear function, this layer does not include a bias term and activation function. Sparse subspace clustering is realized by assuming sparsity and zero diagonal on Γ (Elhamifar and Vidal, 2013):

$$Z_{\text{dec}} \cong Z_{\text{enc}}\Gamma \qquad \text{such that:} \qquad \text{diag}(\Gamma) = 0 \tag{4}$$

In the above equation, Z_{enc} and Z_{dec} are matrices that contain z^i_{enc} and z^i_{dec} for all training images as their columns. This formulation will enforce that the representation of the i^{th} image, z^i_{dec}, be linearly approximated by a small number of those of other images in the training set. Even though the relation between Z_{dec} and Z_{enc} is linear, the clustering method is, in general, very complex because z^i_{enc} is a highly non-linear representation of the image.

To improve the training speed and stability, we first train a standard CAE, i.e., with $\Gamma = I$. We train this standard CAE using the standard CAE cost function, which is the reconstruction error $\|\hat{X} - X\|_2^2$. Here X and $\hat{X}$ are matrices that contain, respectively, the input images and the reconstructed images. After training the standard CAE, we introduce Γ to obtain our full model. We train this model using a cost function that consists of the reconstruction error term and terms for sparse subspace clustering, as shown below:

$$\begin{aligned} &\text{minimize } \|\hat{X} - X\|_2^2 + \lambda_1 \|Z_{\text{enc}} - Z_{\text{enc}}\Gamma\|_2^2 + \lambda_2 \|\Gamma\|_1 \\ &\text{such that } \quad \text{diag}(\Gamma) = 0 \end{aligned} \tag{5}$$

Therefore, in this stage the low-dimensional representations of the images and their similarities are learned jointly. We empirically chose $\lambda_1 = \lambda_2 = 0.1$. For both training stages, we trained the network for 100 epochs using Adam (Kingma and Ba, 2014) with a learning rate of 10^{-3}. After training, an affinity matrix can be created as $C = |\Gamma| + |\Gamma^{\text{T}}|$, where $C(i, j)$ indicates the similarity between the i^{th} and j^{th} images. Here, $|.|$ denotes element-wise absolute value; that is, $|\Gamma|$ indicates a matrix whose elements are absolute values of the elements of Γ. Spectral clustering methods can be used to cluster the training images based on C, but we will use C directly as we explain in Section 2.2.3.

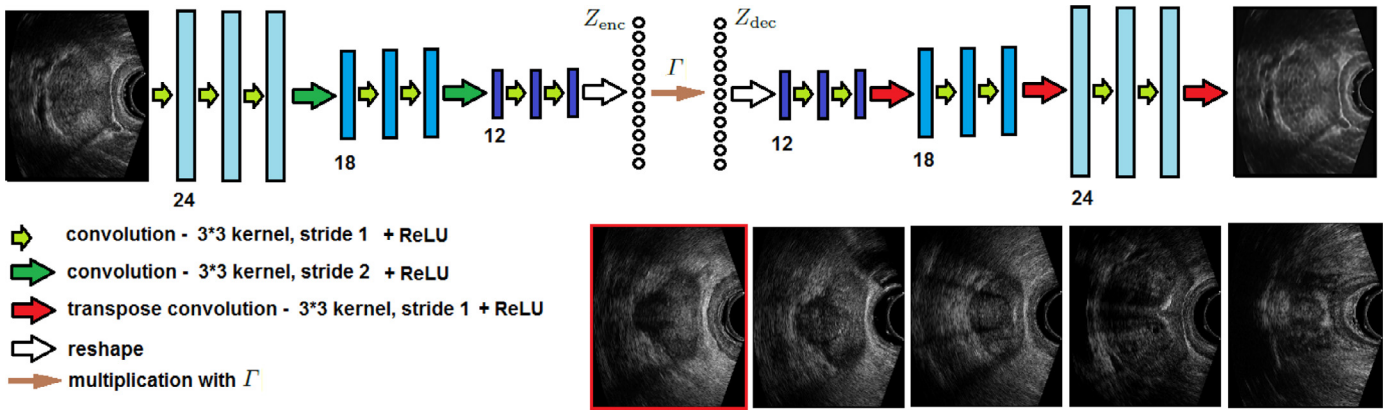

Fig. 2. The CAE architecture used to learn image affinities. On the bottom right, an image (with red borders) is shown along with 4 images with decreasing (left-to-right) similarity to it based on the affinity matrix, $C = |\Gamma| + |\Gamma^T|$, learned by the CAE. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

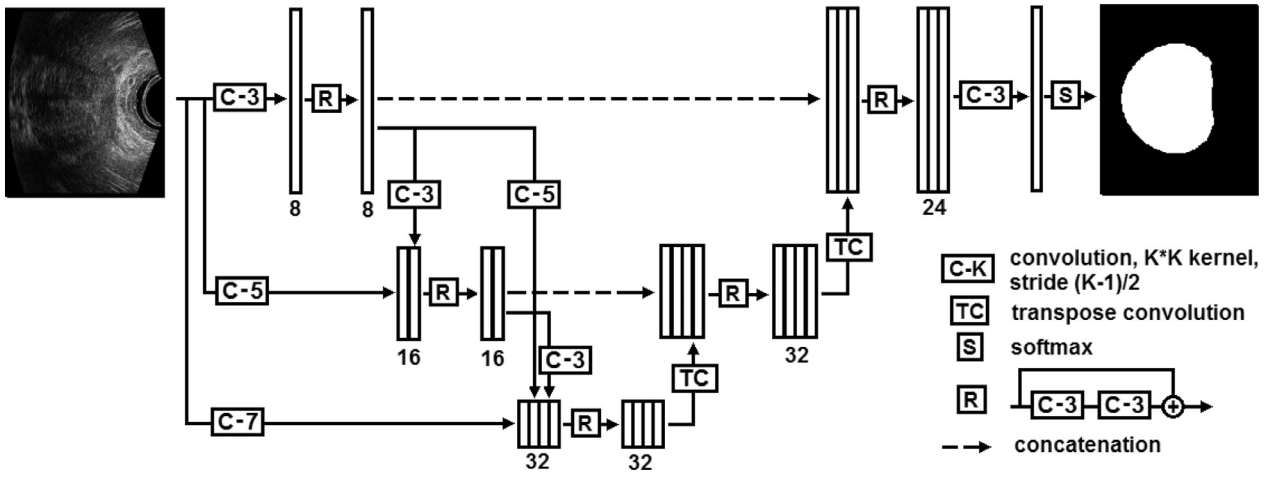

Fig. 3. The proposed CNN architecture. To avoid clutter, the network is shown for a depth of 3. We used a network with a depth of 5; i.e., we also applied C-9 and C-11. Number of feature maps is also shown. All convolutions are followed by leaky ReLU.

2.2.2. Proposed CNN architecture

Our CNN is shown in Fig. 3. Overall, its architecture is similar to the U-Net, proposed by Ronneberger et al. (2015)). However, we modify the U-Net architecture for this specific application in three ways: 1) Whereas U-Net uses the same filter sizes, we apply convolutional filters of different sizes ($k \in \{3, 5, 7, 9, 11\}$) and corresponding strides ($s \in \{1, 2, 3, 4, 5\}$) to the input image. This will help extract fine and coarse features directly from the source image. Moreover, small patches of ultrasound images are overwhelmed by speckle and contain little edge information. Hence, larger filters should help the network learn more informative features at different scales, 2) The finer features are forwarded to all coarser layers after being resized via convolutional kernels of proper sizes and strides. As promoted by the Dense-Net architecture, this improves feature reuse and reduces the number of network parameters (Huang et al., 2017). Therefore, the proposed network extracts features at several different resolutions and fields-of-view. These features are then combined via a series of transpose convolutions to arrive at the final segmentation. 3) Unlike U-Net, all features are passed through residual blocks. This will increase the feature richness ease the training.

The final network layer is passed through a softmax layer to output a segmentation probability map (in [0,1]). The network is trained by maximizing the DSC between this probability map and the ground-truth segmentation. For this, we used Adam with a learning rate of 10^{-4} and performed 200 epochs. The training process is explained in more detail below.

2.2.3. Training a CNN ensemble with adaptive sampling

The loss surface of deep neural networks are non-convex and extremely complex. As a result, deep CNNs converge to a local critical point (Choromanska et al., 2015). When the training data is small, the obtained solution of this optimization procedure can be heavily influenced by the more prevalent training samples. An effective approach to reducing the sensitivity to local minima and improving the generalization accuracy is to learn an ensemble of models (Goodfellow et al., 2016). In this study, we trained $K = 5$ CNN models using 5-fold cross validation. We denote the indices of the training and validation images with S_{tr} and S_{vl}, respectively. Let us also use e_i to denote the "error" committed on the i^{th} validation image by the CNN after the current training epoch. As shown in Fig. 4, for the next epoch we sample the training images according to their similarity to the difficult images in the validation set. Specifically, we sample the j^{th} training image with a probability computed as follows:

$$p(j) = q(j)/\Sigma_j q(j) \quad \text{where} \quad q(j) = \Sigma_{i \in S_{vl}} C(i, j) e(i). \tag{6}$$

At the beginning of training, we initialize p to a uniform distribution. It is important to note that we have great freedom in the choice of the error, e. For example, e does not have to be differentiable. Because one of our aims is to reduce large segmentation

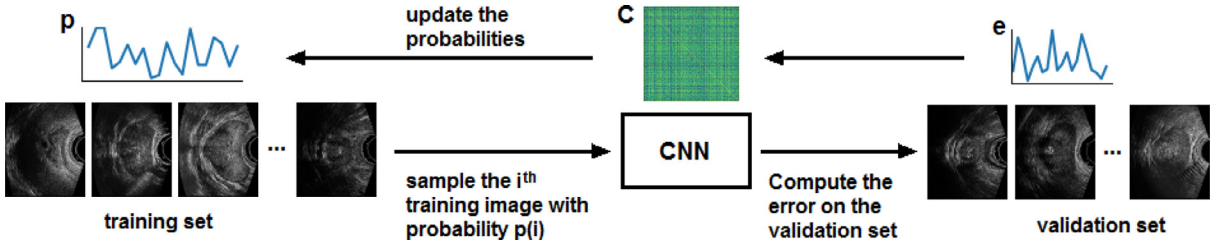

Fig. 4. The proposed training loop with adaptive sampling of the training images.

errors on difficult images, we chose e to be the Hausdorff Distance (HD). For two point sets, X and Y, HD is defined as:

$$\mathrm{HD}(X,Y) = \max\Big(\sup_{y\in Y}\inf_{x\in X}\|x-y\|, \sup_{x\in X}\inf_{y\in Y}\|x-y\|\Big). \quad (7)$$

In image segmentation, X and Y will be the boundaries of the ground-truth and predicted segmentations, which consist of curves in 2D and surfaces in 3D. Although HD is an important measure of segmentation quality, it cannot be easily minimized. Our proposed approach provides an indirect way to reduce HD.

2.2.4. Improving uncertain segmentations using an SSM

As we will explain below, we can estimate the level of confidence in the segmentations by examining the disagreement among the models. We compute the average pair-wise DSC between the segmentations produced by the five CNNs as a measure of agreement among them. If this value is above the empirically-chosen threshold of 0.95, we trust the CNN segmentations because of the high agreement. In such a case, we compute the average of the probability maps produced by the CNN ensemble and threshold it at 0.50 to obtain the final segmentation. An example of this situation is shown in the top row of Fig. 5.

If the estimated agreement is below 0.95, we classify the image as one that is difficult to segment. This can happen for various reasons. For example, an image can be out-of-distribution with regard to the training set, or it can contain strong or unique artifacts. We propose to improve the segmentation of such images by utilizing the prior information in the form of an SSM. In our application, there exist at least two ways of building an SSM. One way is to use the gold-standard CTV segmentations of the training data. The other way, given closeness of CTV and the prostate boundary, is to use accurate segmentation of the prostate in an MRI dataset. We experimented with both options in this work. The modes of shape variation found from the two datasets were very similar. Also, the final segmentation results of difficult TRUS images, which was the focus of this work, were very close for both SSMs. In this paper, we report the results obtained using the SSM from MRI. We built this SSM from a set of 75 MR images with ground-truth prostate segmentation provided by expert radiologists. From each slice of the MR images, we extracted 100 equally-spaced points on the boundary of the prostate. This was done by first computing the length (L) of the boundary and then traversing the boundary from an arbitrary starting point and placing a point when d^*100/L exceeded the next integer, where d is the traversed distance from the start point. All boundary point sets were rigidly (i.e., translation, scale, and rotation) registered to one reference point set. This was followed by deformable registration of the reference point set to the other point sets to determine the point correspondences. Finally, PCA is used to compute the shape variation modes. We built three separate SSMs for base, mid-gland, and apex. In deciding whether an MRI slice belonged to base, mid-gland, or apex, we assumed that each of these three sections accounted for one third of the prostate length. We use u and V to denote, respectively, the mean shape and the matrix with the n most important shape modes as its columns. We chose $n = 7$ because the first 7 modes explained more than 99% of the shape variance.

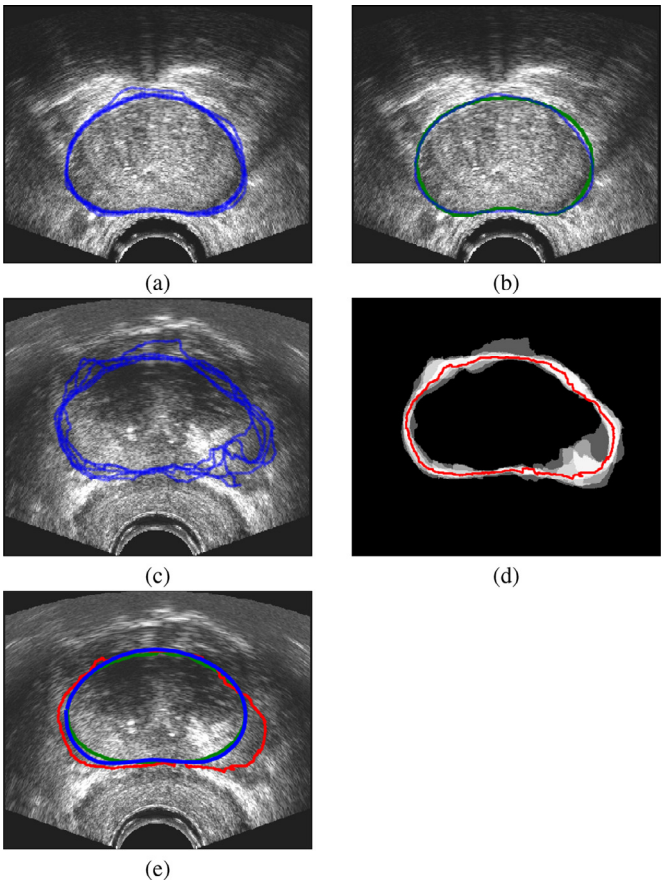

Fig. 5. Top row: an "easy" image, (a) the five CNNs trained on different subsets of the training data produce similar results, (b) the final segmentation produced by thresholding the mean probability map. Bottom two rows: a "difficult" image, (c) there is large disagreement between the five CNN segmentations, (d) the segmentation uncertainty map with s_{init} (red) superimposed, (e) the final segmentation, s_{impr} (blue), obtained using SSM fitting. The green contour shows the ground truth segmentation.. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

If the agreement among the CNN segmentations is below the threshold, we use them to compute a rough initial segmentation boundary as well as a map of segmentation uncertainty. The initial segmentation boundary, denoted with s_{init}, is computed by thresholding the average of the five probability maps, $\bar{p}$, at 0.5. The segmentation uncertainty map is computed as:

$$Q = 1 - \bar{p}^2 - (1-\bar{p})^2 \quad (8)$$

This formulation is based on the Kohavi-Wolpert variance (Kohavi et al., 1996). Q is 0 where all models predict the same class and increases as the certainty in the average ensemble prediction decreases. When all models predict class 0 or class 1

($\bar{p} = 0$ or 1, respectively) Q has its minimum value of 0. On the other hand, when the average predicted probability is $\bar{p} = 0.5$, which corresponds to the highest uncertainty in the average probability prediction, Q will have its maximum value of 0.50. The Kohavi-Wolpert variance has been used to quantify the diversity of classifier ensembles (Kuncheva and Whitaker, 2003). We propose to use it as a measure of classifier (dis)agreement and, hence, (un)certainty.

Using the initial segmentation and the segmentation certainty map, s_{init} and Q, computed as explained above, we estimate an improved segmentation boundary, s_{impr}, as follows:

$$\begin{aligned} s_{\text{impr}} &= R_{\theta^*}[s^*(Vw^* + u)] + t^* \\ &\text{where: } \{s^*, t^*, w^*, \theta^*\} = \\ &\underset{s,t,w,\theta}{\text{argmax}} \frac{2\sum_i p^i_{\text{impr}} p^i_{\text{init}}(1 - Q^i)}{\sum_i (p^i_{\text{impr}})^2(1 - Q^i) + \sum_i (p^i_{\text{init}})^2(1 - Q^i)} \end{aligned} \tag{9}$$

where t, s, and w denote, respectively, translation, scale, and the coefficients of the shape model, R_θ is the planar rotation matrix with angle θ, and p_{init} and p_{impr} denote, respectively, the binary segmentations representing the interior of s_{init} and s_{impr}. The index i in the above equation runs over image pixels. The expression that we are maximizing is a modification of the DSC formulation proposed by Milletari et al. (2016), where we take into account the certainty of the initial segmentation. In other words, we fit an SSM to the initial segmentation while attaching more importance to parts of it that have higher certainty.

Since the objective function in Eq. 9 is non-convex, we use alternating minimization to find a stationary point (Cootes et al., 1995). We initialize t to the centroid of the initial segmentation, s to 1, and w and θ to zeros and perform alternating minimization until the objective function reduces by less than 1% in an iteration. Approximately 3 to 5 iterations sufficed to converge to a good result. An example is shown in Fig. 5(c)-(e). More results will be shown in the next section.

Our proposed method for improving the uncertain segmentations using a shape model (Eqs. (8) and (9)) are slightly different than those in our conference paper (Karimi et al., 2018). Although it is based on the same ideas, this new formulation leads to slightly better results. Moreover, it allows us to compare our method of estimating segmentation uncertainty with other methods in an unbiased manner.

Our approach to estimating the segmentation uncertainty is based on the disagreement among an ensemble of CNNs trained on different subsets of the training data. We compare our approach with Epistemic-Aleatoric and Platt scaling methods. We apply both Epistemic-Aleatoric and Platt scaling methods on the same CNN architecture that we used with our proposed method (Fig. 3). For Epistemic-Aleatoric method, the only modification to the network is the addition of a separate head to predict σ. As required by this method, we train the model while using dropout regularization (with $p = 0.20$). The same dropout rate is used during test. For the Platt scaling method, after training the CNN we learn the parameters of the Platt scaling. Each of these methods, similar to our proposed method, predicts a segmentation as well as an uncertainty map. We compare these methods in two ways:

1. We compare the estimated segmentation uncertainty maps in terms of the Expected Calibration Error (ECE) (Naeini et al., 2015; Guo et al., 2017). ECE is computed as:

$$ECE = \sum_{k=1}^{K} P_k |o_k - e_k| \tag{10}$$

This method of computing the ECE is based on dividing the probability range into K bins. P_k denotes the empirical probability of instances falling into bin k. o_k and e_k denote, respectively, the true fraction of positive instances and the average of the predicted probability of the instances in that bin. Clearly, a lower ECE indicates a better-calibrated model.

2. In terms of segmentation accuracy. For each method, we fit our SSM to the predicted segmentation boundary and the segmentation uncertainty map using the method described in Section 2.2.4 and evaluate the segmentation accuracy.

3. Results and discussion

In the first part of this section, we compare our proposed segmentation method with several other segmentation methods. Then, in the second part of this section we will study some aspects of our proposed segmentation method, such as our estimation of segmentation uncertainty, through comparisons with alternative approaches.

We compare our segmentation method with three existing methods. The first of these is the adaptive shape model-based method of Yan et al. (2011). This method is based on deformable shape models. However, unlike most other methods that learn the shape model from a set of training images prior to segmentation of a test patient's image, in this method the shape model is updated adaptively during segmentation of the patient's image. We refer to this method as ADSM. We will also compare with two CNN-based methods. The first is the widely-used U-Net architecture proposed by Ronneberger et al. (2015). The second is a more elaborate network proposed by Wang et al. (2018). We will refer to this method as DAF because it relies on deep attentional features (DAF) to better integrate multi-scale features so that genuine information that is relevant for prostate segmentation is amplified while suppressing noisy and irrelevant features.

For our proposed method, we report three sets of results in order to assess the effects of different modules in our segmentation pipeline on the segmentation accuracy: 1) Proposed-OneCNN: We train only one CNN, 2) Proposed-Ensemble: We train five CNNs as explained in Section 2.2.3; on a test image, the final segmentation is obtained by thresholding the average of the five probability maps at 0.5, and 3) Proposed-Full: This represents our complete segmentation method. It identifies and improves uncertain segmentations produced by Proposed-Ensemble with the help of the SSM as explained in Section 2.2.4.

As we mentioned above, our dataset consisted of TRUS images of 675 patients. To make the best use of our data, we performed five-fold cross-validation, each time training the model on images from 540 patients and testing on the remaining 135 patients. In this section, we use DSC and HD to quantify the segmentation performance. We also report the 95th percentile of HD across the test images as a measure of the worst-case performance on the population of test images.

The results of this comparison have been summarized in Table 1. Our method outperformed the other methods in terms of DSC and HD. Paired t-tests (at $p = 0.01$) showed that both DSC and HD obtained by our method, Proposed-Full, were significantly better than the other three methods in all three prostate sections. Our method also achieved much smaller values for the 95th percentile of HD. As expected, compared with ADSM that is based on deformable models, our method and the other two deep learning-based methods (U-Net and DAF) achieved better results in terms of DSC. However, neither U-Net nor DAF achieved much better results than ADSM in terms of the measures of worst error, i.e., HD and especially the 95th percentile of HD. This suggests that even though CNN-based methods can achieve quite acceptable results on typical images, on more difficult images their error can be substantial. Fig. 6 shows example segmentations produced by different methods.

Table 2 shows the effectiveness of our proposed strategies for improving the CNN segmentations. Compared with

Table 1
Summary of the comparison of the proposed method with ADSM, U-Net, and DAF.

		DSC	HD (mm)	95th percentile of HD (mm)
Mid-gland	ADSM	87.8 ± 5.9	3.6 ± 2.1	8.0
	U-NET	90.5 ± 3.8	3.7 ± 2.3	7.4
	DAF	92.1 ± 3.8	3.5 ± 2.5	7.3
	Proposed-Full	94.4 ± 3.4	2.5 ± 1.7	4.9
Base	ADSM	85.0 ± 7.1	4.9 ± 3.4	8.6
	U-NET	89.6 ± 4.0	3.8 ± 3.0	8.5
	DAF	91.4 ± 4.3	3.9 ± 2.8	8.8
	Proposed-Full	93.0 ± 3.6	2.5 ± 2.5	5.2
Apex	ADSM	84.3 ± 8.0	4.8 ± 3.2	9.0
	U-NET	86.8 ± 5.7	4.6 ± 3.3	8.8
	DAF	88.5 ± 5.2	4.2 ± 2.9	8.0
	Proposed-Full	91.0 ± 4.9	3.1 ± 1.8	5.5

Table 2
Performance of the proposed method at different stages. Different superscripts (*, **, and ***) in DSC and HD columns indicate statistical difference at $p = 0.01$.

	DSC	HD (mm)	95th percentile of HD (mm)
Proposed-OneCNN	$91.9 \pm 4.6^*$	$3.6 \pm 2.6^*$	8.8
Proposed-Ensemble	$93.7 \pm 3.8^{**}$	$3.2 \pm 2.3^{**}$	6.2
Proposed-Full	$93.9 \pm 3.5^{**}$	$2.7 \pm 2.3^{***}$	5.4

Proposed-OneCNN, Proposed-Ensemble and Proposed-Full achieve much better results in terms of both DSC and HD. Our proposed strategies for dealing with difficult images have greatly reduced the mean, standard deviation, and the 95th percentile of HD. Paired t-tests (at $p = .01$) indicated that Proposed-Ensemble achieved a significantly better DSC and HD than Proposed-OneCNN.

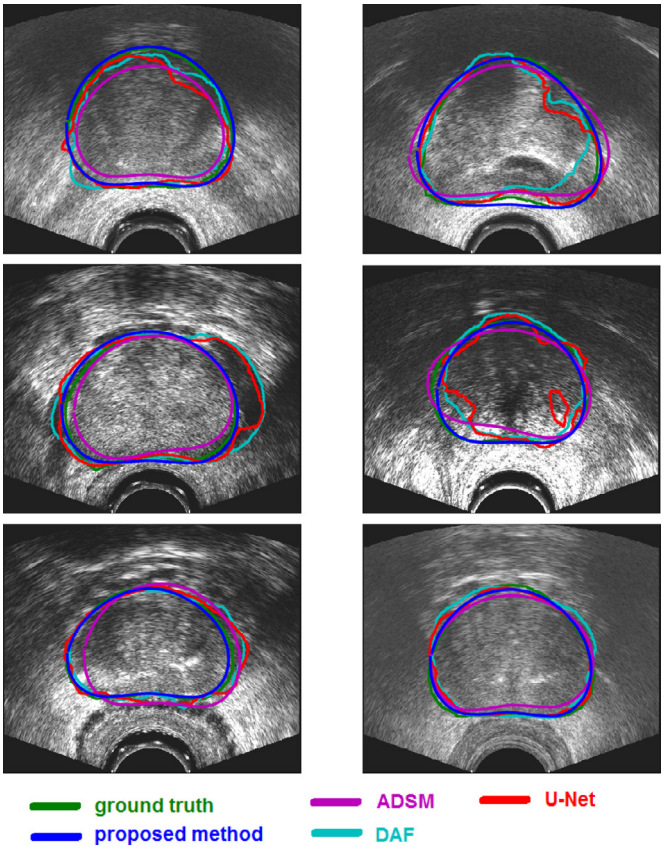

Fig. 6. Example segmentations produced by different methods.

Table 3
Comparison of different approaches to segmentation uncertainty estimation in terms of the accuracy of the final segmentation after SSM fitting.

	DSC	HD (mm)	95th percentile of HD (mm)
Proposed method	93.9 ± 3.5	2.7 ± 2.3	5.4
Epistemic-Aleatoric method	93.1 ± 3.6	3.0 ± 2.1	5.4
Platt scaling method	91.2 ± 4.0	3.1 ± 2.7	6.5
STAPLE	93.2 ± 3.0	3.5 ± 2.7	7.9

Similarly, on images that went through SSM fitting, Proposed-Full significantly improved HD compared with Proposed-Ensemble.

For the proposed method, ECE was 0.0272 ± 0.010. For the Epistemic-Aleatoric and Platt scaling methods, this value was 0.0260 ± 0.010 and 0.0301 ± 0.019, respectively. All three methods demonstrate a very small ECE. This is because for most images they can accurately and with high confidence identify the background and a large part of the foreground. Nonetheless, the Platt method, which is the simplest of the three methods produces more poorly-calibrated segmentation predictions. The Epistemic-Aleatoric method estimates the prediction uncertainty slightly better than our method based on ECE.

A more practically useful comparison of these uncertainty estimation methods is their effect on the final segmentation. Table 3 shows a summary of the segmentation accuracy obtained by applying the SSM fitting to the initial segmentation and the uncertainty maps produced by the three methods. In addition, we have included the results of applying the STAPLE algorithm on the five segmentations produced by the CNN ensemble in our method. In Fig. 7, we have shown two example images with the results of these algorithms.

Overall, the proposed method and the Epistemic-Aleatoric method achieve comparable results, which are better than both Platt scaling method and STAPLE. Often, the segmentation uncertainty map produced by the proposed method and epistemic uncertainty map produced by the Epistemic-Aleatoric method have clear similarities, even though they are based on quite different definitions of the uncertainty and are estimated using very different methods. The segmentation uncertainty map estimated by Platt scaling method, on the other hand, usually looks quite different than those estimated by the proposed method and the Epistemic-Aleatoric method. The aleatoric uncertainty map estimated by the Epistemic-Aleatoric method looks quite uninformative and usually has large values only near the image borders. It has been shown that in computer vision applications, in general, epistemic and aleatoric certainties are more relevant for small and large training datasets, respectively (Kendall and Gal, 2017). Compared to typical medical image analysis applications, the amount of training data available in our study is large. Nonetheless, our results indicate that the epistemic uncertainty is much more significant than the aleatoric uncertainty. In fact, on average epistemic uncertainty was an order of magnitude larger than aleatoric uncertainty. This indicates that, at least in this specific application, and perhaps in CNN-based medical image segmentation in general, the uncertainty due to the small size of training data is more significant than that due to observation noise.

The STAPLE algorithm, which analyzes the agreement level among multiple segmentations without taking into account the expected shape variations, achieved high DSC but much poorer HD than the proposed method. This observation, too, underscores the importance of exploiting the prior shape information in this application. As can be seen in the examples in Fig. 7, on some difficult test images all CNNs in the ensemble can commit the same segmentation error, which will be impossible for an expectation-maximization-based method to fix.

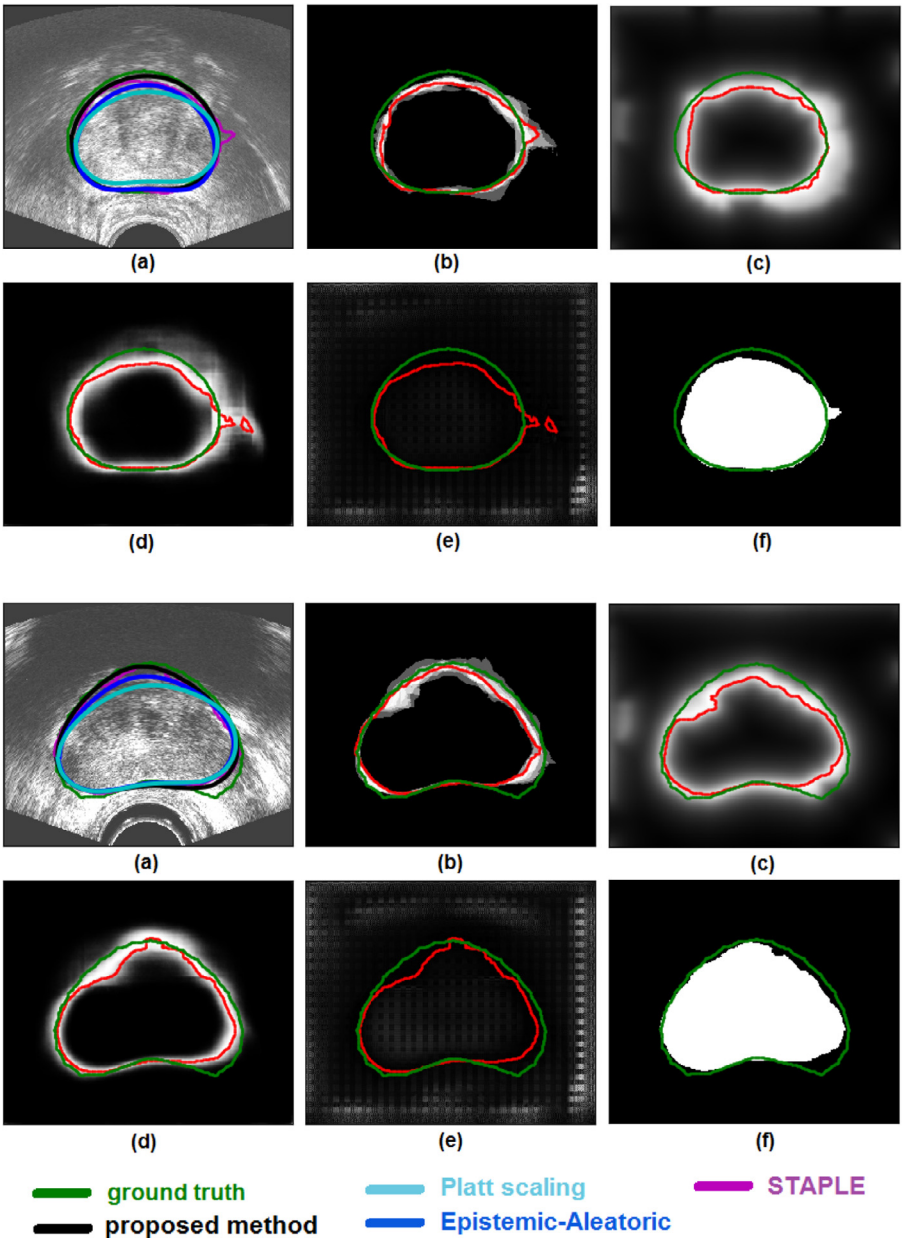

Fig. 7. Two example images comparing the results of different methods for estimating segmentation uncertainty. In each of the two figures: (a) the image along with the final segmentation obtained with different methods, (b) the segmentation uncertainty map computed using the proposed method, (c) the same for Platt scaling method, (d) and (e), respectively, the epistemic and aleatoric segmentation uncertainties estimated by Epistemic-Aleatoric method, and (f) the segmentation obtained with STAPLE algorithm. In all images, the green curve shows the ground-truth and the red curve shows the computed segmentation. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

The results presented in this paper suffer from certain limitations. All images used in this study were collected using ultrasound machines and probe from a single manufacturer. Although we believe our methods should be applicable to other imaging systems, some of the settings may need further tuning when applied on images collected using different systems. As an example, we used a threshold of 0.95 on the mean pair-wise DSC between the segmentations produced by the CNNs to decide if the segmentation was reliable. This value may not be a good setting for images that look very different, and a better setting may be selected empirically based on the range of uncertainties in the new data.

Another limitation of this work was the accuracy of the ground-truth segmentations used for developing our models and for testing them. As we mentioned above, intra- and inter-observer variability for this task were evaluated by Mahdavi et al. (2011). One of the criteria used in that evaluation was the “volume error”, which is equal to 1-DSC. It was observed that the average intra- and inter-observer variability in terms of volume error for manual segmentation were 6.0% and 4.6%, respectively. When using the semi-automatic software, these values were 3.5% and 3.0%, respectively. These values correspond to DSC values in the range between 94.0 and 97.0. In comparison, the highest DSC values achieved in this study were approximately between 93.0 and 94.0, which is close to the range of intra- and inter-observer variability. Given that the inter-observer variability creates an unavoidable error in the ground-truth, we think that this comparison shows that the accuracies achieved by our proposed method are reasonably acceptable for clinical applications.

Both the CAE (Fig. 2) and the CNN (Fig. 3) were implemented in TensorFlow. On an Nvidia GeForce GTX TITAN X GPU, the training times for the CAE and each of the CNNs, respectively, were approximately 24 and 12 hours. For a test image, each CNN produces a

segmentation in 0.02 second. This indicates that the CTV of a patient on approximately 10 images can be computed in a fraction of second. Therefore, our proposed method is well-suited for real-time segmentation.

4. Conclusion

We proposed adaptive sampling of the training data, ensemble learning, and use of prior shape information to improve the accuracy and robustness of prostate CTV segmentation in TRUS images and to reduce the likelihood of committing large segmentation errors. Compared to other traditional and CNN-based methods, our method achieved significantly better results in terms of HD, which measures largest segmentation error. Moreover, our method also substantially reduced the maximum errors on the population of test images. Our results show that even for TRUS images that are noisier than many other medical imaging modalities, the aleatoric uncertainty due to observation noise is far less significant than the uncertainty caused by the small size of training data. Our results further showed that effectively combining prior shape information with segmentation uncertainty can lead to more accurate segmentations than using a probabilistic framework such as STAPLE.

In this work, we used segmentation uncertainty maps to improve the segmentation accuracy. Such uncertainty maps can have other useful applications, such as in registration of TRUS to pre-operative MRI and for radiation treatment planning. A shortcoming of this work is with regard to our ground-truth segmentations, which have been provided by expert radiation oncologists on TRUS images. These segmentations can be biased at the prostate base and apex. A more accurate comparison with registered MRI is, therefore, warranted.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was supported by Prostate Cancer Canada, the Canadian Institute of Health Research, the Natural Sciences and Engineering Research Council, and the C.A. Laszlo Chair in Biomedical Engineering held by S. Salcudean.

References

Anas, E.M.A., Nouranian, S., Mahdavi, S.S., Spadinger, I., Morris, W.J., Salcudean, S.E., Mousavi, P., Abolmaesumi, P., 2017. Clinical target-volume delineation in prostate brachytherapy using residual neural networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 365–373.

Barra, V., Boire, J.-Y., 2001. Automatic segmentation of subcortical brain structures in mr images using information fusion. IEEE Trans. Med. Imag. 20 (7), 549–558.

Betrouni, N., Vermandel, M., Pasquier, D., Maouche, S., Rousseau, J., 2005. Segmentation of abdominal ultrasound images of the prostate using a priori information and an adapted noise filter. Comput. Med. Imag. Graph. 29 (1), 43–51.

Blake, A., Curwen, R., Zisserman, A., 1993. A framework for spatiotemporal control in the tracking of visual contours. Int. J. Comput. Vis. 11 (2), 127–145.

Choromanska, A., Henaff, M., Mathieu, M., Arous, G.B., LeCun, Y., 2015. The loss surfaces of multilayer networks. In: Artificial Intelligence and Statistics, pp. 192–204.

Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J., 1995. Active shape models-their training and application. Comput. Vis. Image Understand. 61 (1), 38–59.

Elhamifar, E., Vidal, R., 2013. Sparse subspace clustering: algorithm, theory, and applications. IEEE Trans. Pattern Anal. Mach. Intell. 35 (11), 2765–2781.

Gal, Y., 2016. Uncertainty in Deep Learning. University of Cambridge.

Gal, Y., Ghahramani, Z., 2015. Bayesian convolutional neural networks with bernoulli approximate variational inference. arXiv: 1506.02158.

Ghavami, N., Hu, Y., Bonmati, E., Rodell, R., Gibson, E., Moore, C., Barratt, D., 2018. Integration of spatial information in convolutional neural networks for automatic segmentation of intraoperative transrectal ultrasound images. J. Med. Imag. 6 (1), 011003.

Ghose, S., Oliver, A., Martí, R., Lladó, X., Vilanova, J.C., Freixenet, J., Mitra, J., Sidibé, D., Meriaudeau, F., 2012. A survey of prostate segmentation methodologies in ultrasound, magnetic resonance and computed tomography images. Comput. Method. Progr. Biomed. 108 (1), 262–287.

Ghose, S., Oliver, A., Mitra, J., Martí, R., Lladó, X., Freixenet, J., Sidibé, D., Vilanova, J.C., Comet, J., Meriaudeau, F., 2013. A supervised learning framework of statistical shape and probability priors for automatic prostate segmentation in ultrasound images. Med. Image Anal. 17 (6), 587–600.

Gibson, E., Hu, Y., Ghavami, N., Ahmed, H.U., Moore, C., Emberton, M., Huisman, H.J., Barratt, D.C., 2018. Inter-site variability in prostate segmentation accuracy using deep learning. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 506–514.

Goodfellow, I., Bengio, Y., Courville, A., Bengio, Y., 2016. Deep Learning, 1. MIT press Cambridge.

Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q., 2017. On calibration of modern neural networks. arXiv: 1706.04599.

Hodge, A.C., Fenster, A., Downey, D.B., Ladak, H.M., 2006. Prostate boundary segmentation from ultrasound images using 2d active shape models: optimisation and extension to 3d. Comput. Method. Progr. Biomed. 84 (2), 99–113.

Huang, G., Liu, Z., Weinberger, K.Q., van der Maaten, L., 2017. Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1, p. 3.

Jemal, A., Bray, F., Center, M.M., Ferlay, J., Ward, E., Forman, D., 2011. Global cancer statistics. CA: A Cancer J. Clin. 61 (2), 69–90.

Jendoubi, A., Zeng, J., Chouikha, M.F., 2004. Segmentation of prostate ultrasound images using an improved snakes model. In: Signal Processing, 2004. Proceedings. ICSP'04. 2004 7th International Conference on, 3. IEEE, pp. 2568–2571.

Ji, P., Zhang, T., Li, H., Salzmann, M., Reid, I., 2017. Deep subspace clustering networks. In: Advances in Neural Information Processing Systems, pp. 23–32.

Kachouie, N.N., Fieguth, P., Rahnamayan, S., 2006. An elliptical level set method for automatic TRUS prostate image segmentation. In: Signal Processing and Information Technology, 2006 IEEE International Symposium on. IEEE, pp. 191–196.

Karimi, D., Zeng, Q., Mathur, P., Avinash, A., Mahdavi, S., Spadinger, I., Abolmaesumi, P., Salcudean, S., 2018. Accurate and robust segmentation of the clinical target volume for prostate brachytherapy. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2018. Springer International Publishing, Cham, pp. 531–539.

Kendall, A., Cipolla, R., 2016. Modelling uncertainty in deep learning for camera relocalization. In: 2016 IEEE International Conference on Robotics and Automation (ICRA). IEEE, pp. 4762–4769.

Kendall, A., Gal, Y., 2017. What uncertainties do we need in Bayesian deep learning for computer vision? In: Advances in Neural Information Processing Systems, pp. 5574–5584.

Kingma, D.P., Ba, J., 2014. Adam: a method for stochastic optimization. In: Proceedings of the 3rd International Conference on Learning Representations (ICLR).

Kohavi, R., Wolpert, D.H., et al., 1996. Bias plus variance decomposition for zero-one loss functions. ICML, 96. 275–83

Kuncheva, L., Whitaker, C., 2003. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. Mach. Learn. 51 (2), 181–207.

Lakshminarayanan, B., Pritzel, A., Blundell, C., 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In: Advances in Neural Information Processing Systems, pp. 6402–6413.

Li, X., Li, C., Fedorov, A., Kapur, T., Yang, X., 2016. Segmentation of prostate from ultrasound images using level sets on active band and intensity variation across edges. Med. Phys. 43 (6Part1), 3090–3103.

Mahdavi, S.S., Chng, N., Spadinger, I., Morris, W.J., Salcudean, S.E., 2011. Semi-automatic segmentation for prostate interventions. Med. Image Anal. 15 (2), 226–237.

Milletari, F., Navab, N., Ahmadi, S.-A., 2016. V-net: fully convolutional neural networks for volumetric medical image segmentation. In: 3D Vision (3DV), 2016 Fourth International Conference on. IEEE, pp. 565–571.

Morris, W.J., Keyes, M., Spadinger, I., Kwan, W., Liu, M., McKenzie, M., Pai, H., Pickles, T., Tyldesley, S., 2013. Population-based 10-year oncologic outcomes after low-dose-rate brachytherapy for low-risk and intermediate-risk prostate cancer. Cancer 119 (8), 1537–1546.

Mozaffari, M.H., Lee, W., 2016. 3d ultrasound image segmentation: a survey. arXiv: 1611.09811.

Naeini, M.P., Cooper, G.F., Hauskrecht, M., 2015. Obtaining well calibrated probabilities using Bayesian binning. In: AAAI, pp. 2901–2907.

Nag, S., Bice, W., DeWyngaert, K., Prestidge, B., Stock, R., Yu, Y., 2000. The american brachytherapy society recommendations for permanent prostate brachytherapy postimplant dosimetric analysis. Int. J. Radiat. Oncol. Biol. Phys. 46 (1), 221–230.

Neal, R.M., 2012. Bayesian Learning for Neural Networks, 118. Springer Science & Business Media.

Noble, J.A., Boukerroui, D., 2006. Ultrasound image segmentation: a survey. IEEE Trans. Med. Imag. 25 (8), 987–1010.

Nouranian, S., Mahdavi, S.S., Spadinger, I., Morris, W.J., Salcudean, S.E., Abolmaesumi, P., 2015. A multi-atlas-based segmentation framework for prostate brachytherapy. IEEE Trans. Med. Imag. 34 (4), 950–961.

Nouranian, S., Ramezani, M., Spadinger, I., Morris, W.J., Salcudean, S.E., Abolmaesumi, P., 2016. Learning-based multi-label segmentation of transrectal ultrasound images for prostate brachytherapy. IEEE Trans. Med. Imag. 35 (3), 921–932.

Pawlowski, N., Brock, A., Lee, M.C., Rajchl, M., Glocker, B., 2017. Implicit weight uncertainty in neural networks. arXiv: 1711.01297.

Qiu, W., Yuan, J., Ukwatta, E., Fenster, A., 2015. Rotationally reslice 3d prostate TRUS segmentation using convex optimization with shape priors. Med. Phys. 42 (2), 877–891.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 234–241.

Salembier, C., Lavagnini, P., Nickers, P., Mangili, P., Rijnders, A., Polo, A., Venselaar, J., Hoskin, P., et al., 2007. Tumour and target volumes in permanent prostate brachytherapy: a supplement to the estro/eau/eortc recommendations on prostate brachytherapy. Radiother. Oncol. 83 (1), 3–10.

Saroul, L., Bernard, O., Vray, D., Friboulet, D., 2008. Prostate segmentation in echographic images: avariational approach using deformable super-ellipse and rayleigh distribution. In: Biomedical Imaging: From Nano to Macro, 2008. ISBI 2008. 5th IEEE International Symposium on. IEEE, pp. 129–132.

Srivastava, N., Hinton, G.E., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 15 (1), 1929–1958.

Sylvester, J.E., Grimm, P.D., Eulau, S.M., Takamiya, R.K., Naidoo, D., 2009. Permanent prostate brachytherapy preplanned technique: the modern seattle method step-by-step and dosimetric outcomes. Brachytherapy 8 (2), 197–206.

Wang, Y., Deng, Z., Hu, X., Zhu, L., Yang, X., Xu, X., Heng, P.-A., Ni, D., 2018. Deep attentional features for prostate segmentation in ultrasound. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 523–530.

Warfield, S.K., Zou, K.H., Wells, W.M., 2004. Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation. IEEE Trans. Med. Imag. 23 (7), 903–921.

Yan, P., Xu, S., Turkbey, B., Kruecker, J., 2011. Adaptively learning local shape statistics for prostate segmentation in ultrasound. IEEE Trans. Biomed. Eng. 58 (3), 633–641.

Yang, X., Fei, B., 2012. 3D prostate segmentation of ultrasound images combining longitudinal image registration and machine learning. In: Medical Imaging 2012: Image-Guided Procedures, Robotic Interventions, and Modeling, 8316. International Society for Optics and Photonics, p. 83162O.

Zaim, A., Jankun, J., 2007. An energy-based segmentation of prostate from ultrasouind images using dot-pattern select cells. In: Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on, 1. IEEE, pp. I–297.

Zeng, Q., Samei, G., Karimi, D., Kesch, C., Mahdavi, S.S., Abolmaesumi, P., Salcudean, S.E., 2018. Prostate segmentation in transrectal ultrasound using magnetic resonance imaging priors. Int. J. Comput. Assist.Radiol. Surg. 1–9.