International Journal of Computer Assisted Radiology and Surgery (2020) 15:1437–1444
https://doi.org/10.1007/s11548-020-02223-x

ORIGINAL ARTICLE

Fast interactive medical image segmentation with weakly supervised deep learning method

Kibrom Berihu Girum[1,2] · Gilles Créhange[1,2] · Raabid Hussain[1] · Alain Lalande[1,3]

Received: 10 January 2020 / Accepted: 20 June 2020 / Published online: 11 July 2020

Abstract

Purpose To achieve accurate image segmentation, which is the first critical step in medical image analysis and interventions, using deep neural networks seems a promising approach provided sufficiently large and diverse annotated data from experts. However, annotated datasets are often limited because it is prone to variations in acquisition parameters and require high-level expert's knowledge, and manually labeling targets by tracing their contour is often laborious. Developing fast, interactive, and weakly supervised deep learning methods is thus highly desirable.

Methods We propose a new efficient deep learning method to accurately segment targets from images while generating an annotated dataset for deep learning methods. It involves a generative neural network-based prior-knowledge prediction from pseudo-contour landmarks. The predicted prior knowledge (i.e., contour proposal) is then refined using a convolutional neural network that leverages the information from the predicted prior knowledge and the raw input image. Our method was evaluated on a clinical database of 145 intraoperative ultrasound and 78 postoperative CT images of image-guided prostate brachytherapy. It was also evaluated on a cardiac multi-structure segmentation from 450 2D echocardiographic images.

Results Experimental results show that our model can segment the prostate clinical target volume in 0.499 s (i.e., 7.79 milliseconds per image) with an average Dice coefficient of 96.9 $\pm$ 0.9% and 95.4 $\pm$ 0.9%, 3D Hausdorff distance of 4.25 $\pm$ 4.58 and 5.17 $\pm$ 1.41 mm, and volumetric overlap ratio of 93.9 $\pm$ 1.80% and 91.3 $\pm$ 1.70 from TRUS and CT images, respectively. It also yielded an average Dice coefficient of 96.3 $\pm$ 1.3% on echocardiographic images.

Conclusions We proposed and evaluated a fast, interactive deep learning method for accurate medical image segmentation. Moreover, our approach has the potential to solve the bottleneck of deep learning methods in adapting to inter-clinical variations and speed up the annotation processes.

Keywords Weakly supervised segmentation · Domain adaptation · CNN · Generative model · Brachytherapy · Echocardiography

Introduction

Accurate and robust medical image segmentation is often profoundly critical in various clinical applications such as medical image analysis and interventions. For example, in radiology, accurate segmentation of structures such as lung, brain, and prostate has become a requisite [1]. It helps to measure the area and volume of structures, which can then be used for tasks such as radiation treatment planning, intervention visualization in image-guided surgery, and registration between different or same imaging modalities [2]. Indeed, clinically analyzing a large population of datasets is often helpful in improving and evaluating the treatment from the patient outcomes. It requires a large and diverse dataset. Although nowadays, there are an increased number of raw medical datasets, manual analysis of these datasets might not be easy. In this regard, computer-aided image analysis has gained interest and also showed promising results [3]. It can be a fully automatic or semiautomatic method. Semiautomatic methods allow expertise to interact with the methods at different levels, while fully automatic methods do not.

Recently, for example, advances in supervised convolutional neural network (CNN)-based fully automatic methods

✉ Kibrom Berihu Girum
kibrom-berihu_girum@etu.u-bourgogne.fr

1 ImViA Laboratory, University of Burgundy, Dijon, France

2 Radiation Oncology Department, CGFL, Batiment I3M, 64b rue sully, 21000 Dijon, France

3 Medical Imaging Department, CHU Dijon, Dijon, France

showed an improvement in computer-assisted diagnostic and therapeutic procedures [4]. However, CNN-based medical image analysis often requires large annotated datasets from different clinical centers and observers with varying acquisition parameters to learn and generalize for any new image case. Unlike in the real-world datasets for semantic image segmentation, annotated medical image datasets are often limited [1]. Data annotation is often prone to variations in acquisition parameters and requires high-level expertise. Manually labeling targets by tracing their contour is also a laborious process. Besides, although the current fully automatic CNN-based approaches have demonstrated promising performances, they have some common limitations. First, they are highly likely to fail in image cases where the testing dataset distribution is different from the annotated training dataset distribution. For example, trained and developed deep learning models from a given specific clinical database distribution might not perform well on other clinical centers and even from the same center with a small change in acquisition parameters. As the neural network weights are learned to segment an image based on supervision from the training dataset, it might not be easy to generalize for any other new image cases where there is a slight difference from the provided training data distributions. This scenario could be solved by manually interacting or retraining the CNN method with new annotated image cases from the new imaging domain. While the first approach is not possible in most currently available end-to-end deep learning methods, the second approach requires considerable time and expert knowledge to get sufficient annotated data, which is often expensive. It is more complicated in image-guided interventions.

Therefore to address such problems, recently artificial intelligent researchers have focused on developing semiautomatic deep learning methods that can allow experts to interact for better accuracy while increasing annotated datasets to improve the accuracy of fully automatic deep learning methods [5–9]. These interactive image segmentation methods can be categorized into two types. Firstly, the pixel-based annotation approach requires user selection of the target and the background based on pixel clicks [6,7]. For example, Maninis et al. [6] incorporated pixel-wise clicks as heat maps into a convolutional neural network-based image segmentation [10]. Secondly, the polygon annotation approach requires an annotator to provide a box around the object of interest [5,11–13]. Castrejon et al. [12] used a recurrent neural network to generate an outline of instance objects in an image from bounding boxes. It showed a significant speedup in the annotation process. Motivated by these promising results, Acuna et al. [11] proposed a reinforcement learning strategy using graph neural networks [14]. To alleviate the problems of polygon recurrent-based image annotation and segmentation methods such as limitations shape, training, and inference time, Ling et al. [5] introduced a graph neural network to predict and correct all vertices of the target simultaneously. Most of these methods were developed for real-world semantic segmentation such as video annotation [9], robotic RGB-D datasets [7], and an instance object segmentation [5,11,12].

Meanwhile, a semiautomatic deep learning approach has been receiving increased attention in the medical domain. Sakinis et al. [8] proposed a semiautomatic segmentation method using a fully convolutional neural networks (U-net) [4]. They have used limited 2D training datasets first to train the U-net architecture. Then, the system allows manual interaction to segment any other new image cases. Rajchl et al. [13] proposed a neural network classifier-based object segmentation from given bounding boxes. They then modified this approach by considering extreme points on the organ's surface [15]. Although these approaches generally showed promising results in speeding up the annotation process of new training datasets, they are not free of limitations. Firstly, most of these methods require several iterations of training and prediction, which could be time-consuming. Secondly, they are computationally expensive. Moreover, though these method's accuracy could be increased by increasing the user clicks, they are very dependent on the intensity information of the raw input images. Consequently, it might not be easy to use these methods for a new imaging modality application (i.e., domain adaptation or transfer learning).

Then, developing an accurate and fast automatized method that requires minimal manual interaction can be beneficial in solving difficult medical image analysis tasks. To achieve this, a semiautomatic deep neural network seems a promising approach which can be used to: 1) increase ground truth data with minimum expert interaction and knowledge in a relatively short time. These semiautomatically generated annotated datasets can, in turn, be used to develop fully automatic deep learning architectures. This approach can also allow retraining a developed fully automatic deep learning method to incorporate or adapt to new medical imaging domains such as on different clinical centers or acquisition protocols (domain adaptation) [16]; 2) improve image segmentation accuracy using only a few good contrasted image pixels by surpassing the need to delineate targets manually by tracing their full contours. It might also allow reducing the often large inter- and intraobserver variations in annotating uncertain image regions.

In this work, we introduce a simple but effective framework to accurately segment predefined structures using a weakly supervised deep learning method. The proposed method employs a generative neural network for anatomical structure prediction from manually selected landmarks, which is then followed by a convolutional neural network for pixel classification. In short, our work has the following main contributions: 1) We developed an efficient interactive method that can be used for accurate automated image segmentation as well as fast image annotation using deep

learning. 2) We designed a method that leverages the prior knowledge of a target and requires fairly small intensity information from the raw input image. This design, as we will show in the experimental result, allows the network to fine-tune for different imaging modalities. We also artificially modeled the errors that can be introduced from selecting the contour landmarks between inter- and intraobservers. 3) We support our claims by an extensive ablation and experimental results on prostate clinical target volume segmentation from TRUS and CT images. Echocardiographic 2D images have also been used to evaluate our method in multi-structure target segmentation. The experimental results on real patient's datasets using both volume and distance metrics reveal that our method can generate accurate image segmentation results in less than 8 milliseconds.

The rest of the paper is organized as follows: we first describe the materials and our method, we then present the experimental results, and finally, we draw our conclusions.

Materials and methods

To evaluate our system, we consider three different applications: prostate segmentation from transrectal ultrasound (TRUS) and computed tomography (CT) images, and cardiac multi-structure segmentation from ultrasound images.

For the prostate segmentation, we collected clinical databases of intraoperative TRUS images as well as postoperative CT images from image-guided prostate brachytherapy. Both imaging modalities are often used to segment clinically meaningful target volumes in radiotherapy, such as the prostate gland. Thus, for this study, we used a clinical database of 78 CT prostate image cases from the anticancer center of Dijon, France. All patients underwent at least a primary permanent prostate brachytherapy with ^{125}I for localized prostate cancer treatment [2]. The in-plane resolution of these CT data varies from 0.4×0.4 mm^2 to 0.58×0.58 mm^2 with a slice thickness between 1.5 mm and 2.5 mm. The acquisition protocol was in helical mode, 120 kVp, 172 mm FOV, and 440 mAs/slice. The TRUS images were acquired from 145 patients who underwent permanent seed implantation under TRUS guidance. The pixel size of each transverse slice was 0.1038×0.1038 mm with a slice thickness of 1 mm. The prostate was manually delineated on both TRUS and CT images by an experienced radiation oncologist. Indeed, these delineation procedures are routinely used in the clinical treatment of image-guided prostate brachytherapy with the help of VariSeed planning software (Varian Medical Systems Inc, Palo Alto, CA).

The TRUS and CT images were resized and center cropped to an image resolution of $256 \times 256 \times 64$, considering the organ at the center of the cropped image (Fig. 1a). We resized the TRUS images into $0.25 \times 0.25 \times 1\, mm^3$ and CT images into $0.5 \times 0.5 \times 1.25\, mm^3$.

Another application of our method is in cardiac multi-structure segmentation on 2D echocardiographic images. For this application, we selected a public dataset [17]. It consists of 450 raw patient datasets with the heart oriented in long-axis orientation and corresponding ground truths. Each of the data has a different image resolution and was resized into 256×256 (there is only one plane). We automatically extracted the four boundary coordinates from the preprocessed mask for both prostate and cardiac applications (top, bottom, left, and right extreme points) as contour pseudo-landmarks. This preprocessing step of the raw input image is shown in Fig. 1a. For each dataset, all exams were acquired on the same plane orientation. However, to adapt the method for any plane orientation, it might require a dataset acquired from all possible target orientations. Examples of selected contour pseudo-landmarks are shown as red points on the input image in Figs. 1b and 2. However, these landmarks are pixel-wise dependent. Thus, to learn the system for any variation in selecting these landmarks, we artificially introduced and modeled errors in the extracted points. It was used to demonstrate the system's performance according to the inter- and intraobserver variations while selecting the pseudo-landmarks. This is further discussed in the "Experimental setup" section.

Proposed method

When experts delineate a given target on an image, they often consider both their prior knowledge of the target and the intensity distribution of the raw input image simultaneously. Similarly, our method involves prior-knowledge prediction from pseudo-landmarks indicated by the user and the intensity distribution information from the raw input image, as shown in Fig. 1b. We model the prior-knowledge prediction using a deep generative neural network (named prior-knowledge generator) [18]. For this, we considered the four pseudo-contour landmarks as an input, i.e., $I^u = (x_i^u, y_i^u)$, where $i = [1, 2, 3, 4]$ are the pseudo-boundary coordinates (x_i and y_i) of the target in a given image u. Moreover, to classify if the given image is with the target organ or not, we used an additional input of s^u. It has a value of either 0 (no organ in a given image) or 1 (presence of the organ in a given image). Then, the prior-knowledge generator is trained to predict a labeled model of $W \times H$, whose pixel $v = (x, y)$ contains a label 1 for the organ and 0 otherwise, from the 1×9 (four pairs of x^u and y^u along with s^u for a given image u) pseudo-landmark coordinates. Here, W and H are the width and height of the image, respectively.

The predicted prior knowledge is then multiplied with the raw input image and can be considered as target attention or region proposal. It is then further merged with the raw

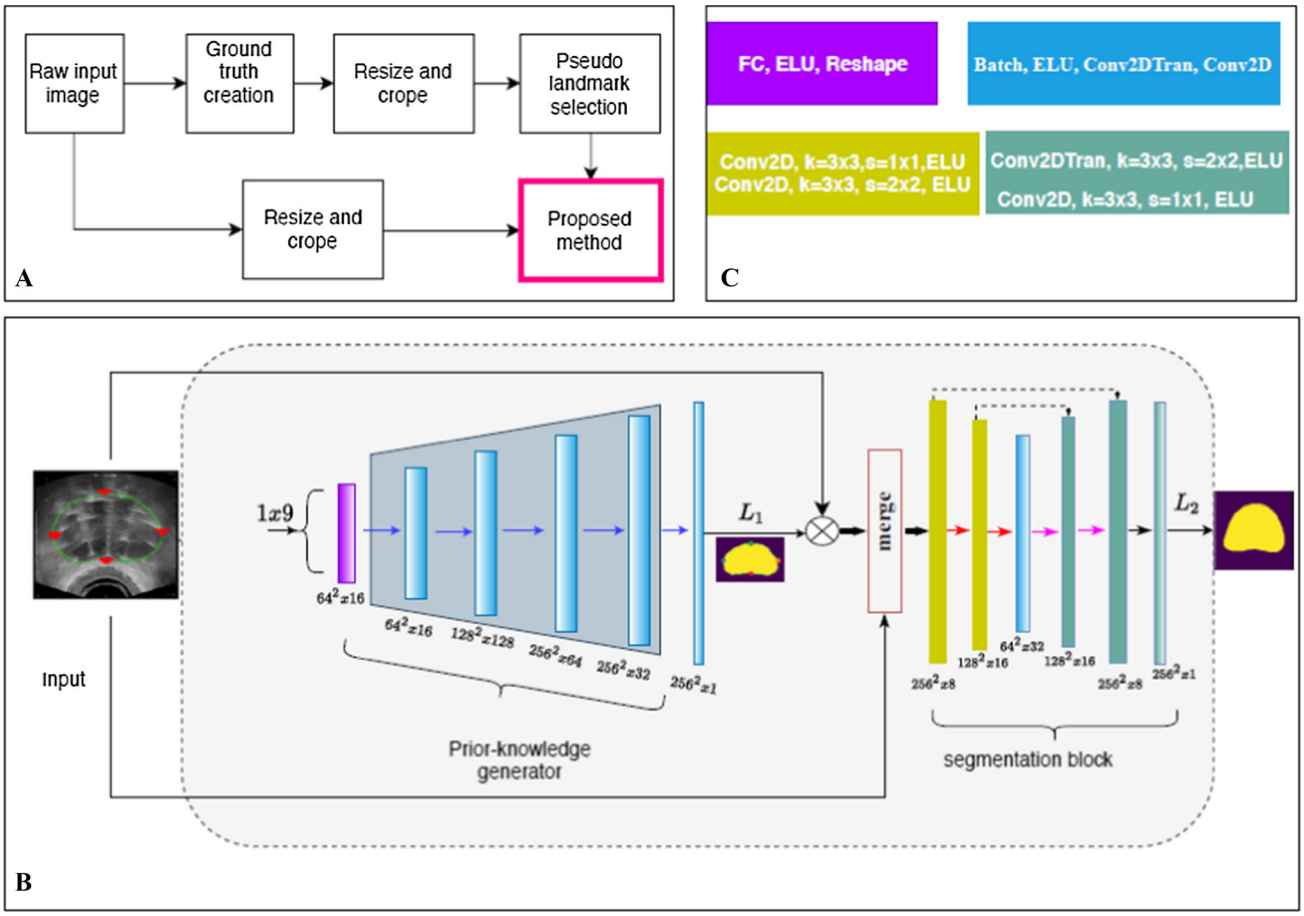

Fig. 1 Proposed method. **a** Preprocessing step; **b** proposed end-to-end architecture; and **c** building blocks of the proposed method, where the top-left bock is for the first layer, and the top-right block is for the rest of the layers of the prior-knowledge generator. Similarly, the bottom-left block is for the down-sampling, and the bottom-right bock is for the up-sampling layers of the segmentation block. The feature channels and image sizes are shown at the bottom of each layer in B

input image using the concatenation layer and fed to the second part of our method, which is a fully convolutional neural network (segmentation block). It is designed to refine further the proposed region from the prior-knowledge generator by leveraging the low-level information from the raw input image.

The complete network architecture is illustrated in Fig. 1. As mentioned before, it consists of two main parts: the prior-knowledge generator and segmentation block. The prior-knowledge generator is composed of a fully connected layer, batch normalization and reshape (Fig. 1c, top-left block), followed by repeated batch normalization, up-convolution and convolution layers (Fig. 1c, top-right block) (similar as in [19]). The fully connected convolutional neural network (segmentation block) is composed of repeated 3×3 convolutions with a stride of 1 and 2, respectively, for the down-sampling (Fig. 1c, bottom-left block). The bottleneck (third block) and the output layer (last block) are convolution layers with stride 1 and kernel size 3×3. In the up-sampling, we used repeated 3x3 deconvolution and convolution layers (Fig 1 c, bottom-right block). In all layers, including the prior-knowledge generator, we used exponential linear unit (ELU) except the last output layers. The last layer's activation function for both the generator and the segmentation block was the same. It was sigmoid for prostate segmentation and softmax for cardiac image segmentation. The number of feature channels in the segmentation block was 8 in the first level, which is then doubled and halved, respectively, after each block of the down-sampling/up-sampling layers. The number of feature channels was doubled at each layer for the cardiac image segmentation.

Training loss function

As the proposed method has two outputs (i.e., prior-knowledge generator's output and final segmentation output), we define a combined loss function as:

$$L_{total} = \frac{1}{2} \times (L_1 + L_2) \quad (1)$$

where L_1 and L_2 are the prior-knowledge generator's loss and final output loss, respectively. We used a sum of binary or categorical cross-entropy and Dice coefficient loss for both L_1 and L_2 (i.e., $L_i = L_{binary} + L_{dice}, i = 1, 2$). The binary cross-entropy was for the prostate segmentation, while the categorical cross-entropy was for the cardiac multi-label segmentation. Although binary/categorical cross-entropy loss function alone is often applied in medical image segmentation, it might not work well for images with small foreground regions. In such cases, the loss function considers all pixels equally. Thus, the small region's information can be suppressed by the larger region's information. It can be solved using the Dice coefficient loss. It is a measure of overlap between predicted segmentation results and reference ground truths. We calculated the Dice coefficient loss, L_{dice}, as:

$$L_{dice} = 1 - \sum_{k \in 0,1} \frac{2 \times \sum_{i \in I} u_i^k v_i^k}{\sum_{i \in I} u_i^k + \sum_{i \in I} v_i^k} \tag{2}$$

where u is the output of the network, and v is the ground truth segmentation map. Both u and v have shape I with $i \in I$ being the number of pixels in the training batch and k being the pixel class. At the output of the prior-knowledge generator and the segmentation block, we adapted a sigmoid function for the 1-channel output in prostate segmentation and a softmax function for the 4-channel output in cardiac image segmentation.

Training, testing configurations, and implementation details

The proposed method was trained using Eq. 1 and an ADAM optimizer. We used a learning rate of 0.0001 and a batch size of 20 until convergence [20]. An early stop of 30 was adapted to determine the convergence. It is important to mention here that we feed the network with 2D, and the predicted image labels are then stacked to create a 3D volume for the prostate segmentation. We considered randomly 25% of each dataset for validation and the remaining 75% for training. The echocardiographic images were also randomly divided into 400 patients for the training and 50 patients for the validation. The proposed network has only 1.5 million trainable parameters, which is fairly very small compared to 32 million trainable parameters in [8]. It was implemented in Python on i7 computer with 32-GB RAM and a dedicated GPU (NVIDIA TITAN X, 12 GB) with Keras API and Tensorflow backend.

Experimental setup

To evaluate the proposed method, we considered both volume and distance metrics (i.e., Dice similarity coefficient (DSC), percent of volume overlap (VO), accuracy (Acc.), and Hausdorff distance (HD)). In all evaluation metrics for the prostate gland segmentation, we consider the 3D volume of the clinical target volume, i.e., each case has a resolution of 256 x 256 x 64. Furthermore, to provide a detailed analysis, the prostate was classified along the transverse axis into 40% for mid-gland and 30% each for base and apex, respectively. Similarly, the echocardiographic image segmentation was evaluated on each structure, such as the left ventricular cavity, the left atrium and the myocardium, and on the whole heart.

Ablation study: To investigate the best combination strategy of the prior-knowledge generator and the raw input image (i.e., merge block in Fig. 1 (B)), we conducted three different combination possibilities such as concatenation, addition, and multiplication. Moreover, we compared the accuracy of our method for the different numbers of landmarks a user has to pinpoint, such as using 2, 4, and 6 pseudo-landmarks.

Inter- and intraobserver variation study: Previous studies reported that the inter- and intraobserver variation in prostate delineation on TRUS images could be in the range of 7.59–27.14% and 8.70%, respectively [21]. In our case, to study the inter- and intraobserver variations that could come while selecting the pseudo-landmarks, we introduced the errors artificially at testing time. We added randomly an error in the range of 1 to 9 mm (i.e., 4- to 36-pixel error) to the selected pseudo-landmarks. Then, the performance of the model in segmenting the prostate gland with such pseudo-landmark error is measured.

Domain adaptation: It is well known that weights of convolutional neural networks are often updated according to the training dataset distribution. However, if the raw input during the testing phase differs from the training data, the model might not work very well. It is because the domain of the input data is changed, while the task domain remains the same. This problem is common in medical image analysis. However, it can be solved using domain adaptation (or transfer learning) [16], in which a trained model on a given dataset could be applied to different domains but for a similar target.

In this study, we conducted ablation experiments in prostate gland segmentation from CT and TRUS images to investigate our method's usage in domain adaptation between different imaging modalities. The shape of the target (i.e., the prostate gland) is the same on both CT and TRUS images. We consider the TRUS data as an available dataset and CT images as the target domain. Then, we performed the two categories of domain adaptation. First, a trained model from TRUS images is applied to CT images, where we take only 20 CT cases to update the trained model (named weakly supervised). Second, we train our model with the TRUS dataset and directly apply it to the testing CT images (named unsupervised). Similarly, we trained the model from CT and applied it to the TRUS images.

Table 1 Quantitative segmentation results. Values are expressed as mean ± std. DSC: Dice coefficient; HD: 3D Hausdorff distance (in mm); and VO: percent of volumetric overlap; Acc: accuracy; TRUS: transrectal ultrasound; CT: computed tomography; US: echocardiographic ultrasound image; LV: left ventricle, MYO: myocardium; LA: left atrium

Image	%	Metric			
		DSC (%)	HD% (mm)	VO (%)	Acc (%)
TRUS	Total	96.9 ± 0.9	4.25 ± 4.58	93.9 ± 1.80	98.9 ± 0.5
	Mid-gland	97.4 ± 0.6	3.28 ± 0.91	94.9 ± 1.13	98.1 ± 0.61
	Apex	96.4 ± 1.0	3.19 ± 1.05	93.1 ± 1.87	98.9 ± 0.36
	Base	96.5 ± 1.5	4.42 ± 4.63	93.4 ± 2.76	98.0 ± 1.14
CT	Total	95.4 ± 0.9	5.17 ± 1.41	91.3 ± 1.70	99.7± 0.11
	Mid-gland	95.9 ± 0.9	4.94 ± 1.62	92.18 ± 1.71	99.3 ± 0.27
	Apex	94.9 ± 1.8	3.56 ± 1.56	90.3 ± 3.20	99.6 ± 0.15
	Base	95.1 ± 1.2	4.92 ± 1.37	90.6 ± 2.21	99.3 ± 0.20
US	Total	96.3 ± 1.3	23.0 ± 10.42	93.3 ± 0.02	98.6 ± 0.01
	LV	93.2 ± 4.9	14.2 ± 5.70	87.2 ± 7.53	98.8 ± 0.01
	LA	91.9 ± 6.6	14.9 ± 5.70	84.9 ± 0.09	99.1 ± 0.01
	MYO	89.5 ± 11.9	22.4 ± 21.08	81.0 ± 0.14	98.0 ± 0.02

Results

The quantitative segmentation results of the proposed method trained and tested from the same modalities (i.e., trained from TRUS and tested on TRUS and similarly on CT images) are shown in Table 1. Although the average Hausdorff distance of the TRUS image is better than the CT image, we observed more variations across the base (i.e., with a standard deviation of 4.63). As shown in Table 1, our method also yielded promising results in cardiac multi-structure segmentation from ultrasound images. However, it appears to produce large Hausdorff distance errors in particular at the level of the epicardial contour of the myocardium. Examples of image segmentation are shown in Fig. 2.

Ablation study: The experimental results for the combination strategy of the prior-knowledge generator and the raw input image are shown in Table 2. We observed that all combination strategies showed a competent segmentation accuracy. However, although multiplication and addition combination strategies can produce competitive results with even smaller parameters, we experimentally observed that they produce more variation in accuracy and high Hausdorff distance error across the testing dataset than the concatenation layer. It is because the addition and multiplication layers influence more the intensity distribution of the original input image before the main segmentation network is seeing it (segmentation block in Fig. 1b). If an error is introduced in the prior-knowledge generator, it would have a high chance to propagate till the end of the network when using addition and multiplication layers than using the concatenation layer. This further suggests that using the concatenation layer would tolerate the inevitable inter- and intraobserver differences in selecting the pseudo-landmarks. Experiments with a different number of landmarks on TRUS image segmentation yielded an average Dice coefficient of 89.7% and

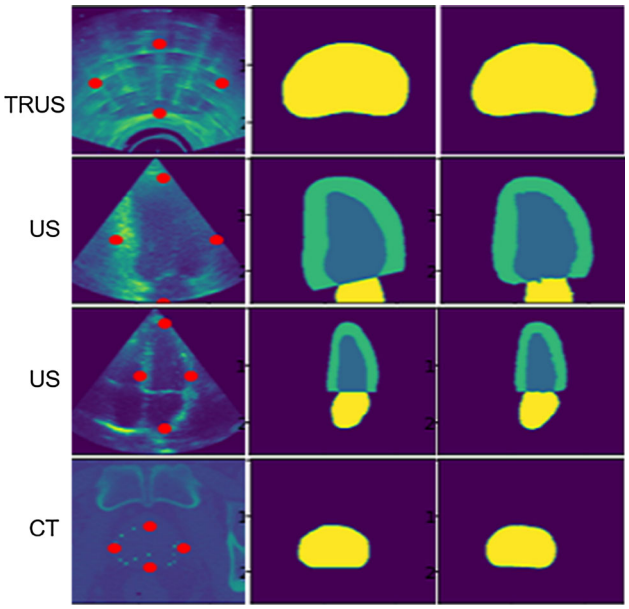

Fig. 2 Segmentation results. The raw input image, ground truth, and segmentation results are shown, respectively, in columns 1, 2, and 3. The four selected pseudo-landmarks are also displayed on the input images with the red dot. TRUS: transrectal ultrasound images, CT: computed tomography images, and US: ultrasound echocardiographic images

97.1%, respectively, for 2 and 6 pseudo-landmarks. The processing time depends on the number of pseudo-landmarks because the user would take time to pinpoint these landmarks. Thus, the more the landmarks, the more time it takes to pinpoint.Then, considering only four landmarks is an excellent compromise.

Inter- and intraobserver variation study: We observed that an error introduced in the user inputs of up to 7 mm does not influence the system both in Hausdorff distance and Dice similarity coefficient. However, user input error of more than 8 mm appears to influence the performance of the system. It is

Table 2 Ablation study of the combination strategies (Concat: concatenation, Add: addition, and Multi: multiplication) on TRUS images

Metric	Concat.	Add.	Mult.
DSC (%)	96.9 ± 0.90	95.9 ± 0.68	96.8 ± 0.67
HD (mm)	4.25 ± 4.58	15.95 ± 8.65	4.94 ± 3.58
VO (%)	93.9 ± 1.80	92.2 ± 1.25	93.8 ± 1.29
Acc (%)	98.9 ± 0.50	98.6 ± 0.44	98.9 ± 0.37

worst in Hausdorff distance. For example, input user errors of 8 mm and 9 mm, respectively, produce a reduced Dice coefficient of 94.2% and 93.8% and an increased Hausdorff distance error of 4.93 and 10.94 mm. However, the uniformity of the introduced errors across the prostate gland's volume would affect the performance; for example, distributed errors among all pseudo-landmarks would be tolerable than a large error introduced on one of the landmarks.

Domain adaptation: As can be seen from Table 3, the trained model from TRUS and retraining using only 20 CT cases can predict the clinical target volume on CT images comparable with using more datasets such as 58 cases (Table 1). Similarly, the trained model from CT and retraining using only 20 TRUS cases yielded a Dice coefficient of 96.5%. It shows that a trained method from a given annotated dataset can be used for other data to segment the same target structure. Moreover, using annotated datasets from magnetic resonance images (MRI) would have been necessary as delineation on these datasets mostly produces less inter- and intraobserver difference. For the unsupervised domain adaptation, the multiplication layer yielded better results than the concatenation layer. It is because, as mentioned in the ablation study section, the prior-knowledge generator suppresses more the input image intensity difference among the modalities while preserving the imaging modality invariant shape of the organ.

For a test exam of size 256 × 256 × 64, our method produces a segmentation result in 0.499 seconds (approximately, 7.79 milliseconds to process an image of size 256 × 256). It was tested on a personal computer of i7 with 32-GB RAM and GPU (GeForce GTX 1070).

The proposed method showed promising results for cross-training with images from different imaging modalities. It can reduce the need to start from zero each time a new modality is required. However, studying and modeling of the challenges and variations in medical imaging acquisitions would be important. Moreover, our approach has shown promising results for the delineation of multi-structure targets from only a few pseudo-landmarks of a particular structure, for example, for cardiac multi-structure segmentation using only four contour landmarks at the level of the myocardium. The Hausdorff distance error, particularly at the myocardium, often with a small area and difficult to segment, could be improved by increasing the depth of the segmentation block. The pseudo-landmarks used in our system to initialize the prior-knowledge prediction could be modeled using a variational auto-encoders or extract automatically from the raw input images.

The datasets used in our experiments were not challenging to resize to the same voxel spacing and center crop because all targets were at the center of the original input image. However, as this may not be the same for other medical imaging domains, it might require high care not to bias the pixel intensity information and not to lose high-level information of the images while interpolating or cropping.

One of the limitations of our method lies in the target structure regularity. As the prior-knowledge generator is designed to learn uniform structures, it would be promising to segment target with uniform shape and topology (such as heart, prostate, liver, or lung). In other segmentation tasks involving irregularly shaped targets such as tumors, it remains unclear how our approach would perform. It will be investigated in future work.

Conclusions

In this paper, we presented a fast, interactive deep learning framework for accurate medical image segmentation and performed extensive ablation studies to apply the system in different imaging domains. It yielded promising segmenta-

Table 3 Domain adaptation study

Training	Testing	Metric	Weakly supervised	Unsupervised	
			Conc.	Conc.	Mult.
TRUS	CT	DSC (%)	94.8 ± 1.71	87.8 ± 6.36	89.1 ± 5.03
		HD (mm)	5.68 ± 2.14	22.9 ± 28.86	10.1 ± 4.28
		VO (%)	90.2 ± 3.03	78.9 ± 9.67	80.7 ± 7.60
		Acc (%)	99.6 ± 0.13	99.1 ± 0.57	99.2 ± 0.42
CT	TRUS	DSC (%)	96.5 ± 0.83	84.7 ± 4.09	84.3 ± 4.02
		HD (mm)	5.20 ± 2.94	28.4 ± 7.78	20.7 ± 6.92
		VO (%)	93.2 ± 1.54	73.6 ± 6.11	73.1 ± 5.97
		Acc (%)	98.8 ± 0.42	95.0 ± 2.00	95.0 ± 1.95

tion results with an average Dice similarity coefficient of 97% to deliver 3D contours on intraoperative transrectal ultrasound images. Similarly, it produced a 95% Dice similarity coefficient on postoperative computed tomography images of prostate brachytherapy. Experiments on cardiac multi-structure segmentation from 2D echocardiographic images also yielded promising results. Without the assumption of a few additional seconds to select the landmarks, our method produces a segmentation result in 0.499 seconds for a test exam of size 256 x 256 x 64. Thus, the proposed method is well suited for real-time prostate clinical target volume segmentation in transrectal ultrasound-based image-guided prostate brachytherapy. We also demonstrated the application of our method for domain adaptation by training on given annotated dataset distribution and applying it onto different testing domains. As it is less dependent on the intensity distribution of the raw input images, it learned to transfer the knowledge of the target between different imaging modalities. This approach could be used to transfer information between different computer-assisted radiology tasks such as in image-guided permanent prostate brachytherapy procedures. We believe that our approach has the potential to solve the bottleneck of deep learning methods in adapting to inter-clinical imaging and dataset variations and speed up the annotation process in weakly supervised-based domain adaption applications.

Acknowledgements The authors would like to thank NVIDIA for providing GPU (NVIDIA TITAN X, 12 GB) through their GPU grant program.

Compliance with ethical standards

Conflicts of interest All authors declare no conflict of interest. Ethical approval and informed consent were not required for this study.

References

1. McBee MP, Awan OA, Colucci AT, Ghobadi CW, Kadom N, Kansagra AP, Tridandapani S, Auffermann WF (2018) Deep learning in radiology. Acad Radiol 25(11):1472–80. https://doi.org/10.1016/j.acra.2018.02.018
2. Girum KB, Lalande A, Quivrin M, Bessières I, Pierrat N, Martin E, Cormier L, Petitfils A, Cosset JM, Créhange G (2018) Inferring postimplant dose distribution of salvage permanent prostate implant (PPI) after primary PPI on CT images. Brachytherapy 17(6):866–73. https://doi.org/10.1016/j.brachy.2018.07.017
3. Litjens G, Kooi T, Bejnordi BE, Setio AA, Ciompi F, Ghafoorian M, van der Laak JA, van Ginneken B, Sánchez CI (2017) A survey on deep learning in medical image analysis. arXiv: 1702.05747
4. Ronneberger O, Fischer P, Brox T (2015) U-net: convolutional networks for biomedical image segmentation. Miccai. https://doi.org/10.1007/978-3-319-24574-4_28
5. ing H, Gao J, Kar A, Chen W, Fidler S (2019) Fast interactive object annotation with curve-gcn. CVPR. 5257–5266. arXiv: 1903.06874
6. Maninis KK, Caelles S, Pont-Tuset J, Van Gool L (2018) Deep extreme cut: From extreme points to object segmentation. CVPR. https://doi.org/10.1109/CVPR.2018.00071
7. Suchi M, Patten T, Fischinger D, Vincze M (2019) EasyLabel: a semi-automatic pixel-wise object annotation tool for creating robotic RGB-D datasets. ICRA. https://doi.org/10.1109/ICRA.2019.8793917
8. Sakinis T, Milletari F, Roth H, Korfiatis P, Kostandy P, Philbrick K, Akkus Z, Xu Z, Xu D, Erickson BJ (2019) Interactive segmentation of medical images through fully convolutional neural networks.1-10. arXiv: 1903.08205
9. Benard A, Gygli M (2017) Interactive video object segmentation in the wild. arXiv: 1801.00269
10. Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille AL (2017) Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE T Pattern Anal. https://doi.org/10.1109/TPAMI.2017.2699184
11. Acuna D, Ling H, Kar A, Fidler S (2018) Efficient interactive annotation of segmentation datasets with polygon-rnn++. CVPR. https://doi.org/10.1109/CVPR.2018.00096
12. Castrejon L, Kundu K, Urtasun R, Fidler S (2017) Annotating object instances with a polygon-rnn. CVPR. https://doi.org/10.1109/CVPR.2017.477
13. Rajchl M, Lee MC, Oktay O, Kamnitsas K, Passerat-Palmbach J, Bai W, Damodaram M, Rutherford MA, Hajnal JV, Kainz B, Rueckert D (2016) Deepcut: object segmentation from bounding box annotations using convolutional neural networks. IEEE T Med Imaging 36(2):674–83. https://doi.org/10.1109/TMI.2016.2621185
14. Li Y, Tarlow D, Brockschmidt M, Zemel R (215) Gated graph sequence neural networks. 1-20. arXiv: 1511.05493
15. Roth H, Zhang L, Yang D, Milletari F, Xu Z, Wang X, Xu D (2019) Weakly supervised segmentation from extreme points. In: Zhou L et al (eds) LABELS 2019, HAL-MICCAI 2019, CuRIOUS 2019. https://doi.org/10.1007/978-3-030-33642-4_5
16. Wang M, Deng W (2018) Deep visual domain adaptation: a survey. Neurocomputing 312:135–53. https://doi.org/10.1016/j.neucom.2018.05.083
17. Leclerc S, Smistad E, Pedrosa J, Østvik A, Cervenansky F, Espinosa F, Espeland T, Berg EA, Jodoin PM, Grenier T, Lartizien C (2019) Deep learning for segmentation using an open large-scale dataset in 2D echocardiography. IEEE T Med Imaging 22 38(9):2198–210. https://doi.org/10.1109/TMI.2019.2900516
18. Radford A, Metz L, Chintala S (2015) Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv: 1511.06434
19. Girum KB, Créhange G, Hussain R, Walker PM, Lalande A (2019) Deep Generative Model-Driven Multimodal Prostate Segmentation. In: Nguyen D, Xing L, Jiang S (eds) Artificial intelligence in radiation therapy. AIRT 2019. https://doi.org/10.1007/978-3-030-32486-5_15
20. Kingma DP, Ba J (2014) Adam: A Method for Stochastic Optimization. 1–15. arXiv: 1412.6980
21. Sandhu GK, Dunscombe P, Meyer T, Pavamani S, Khan R (2012) Inter-and intra-observer variability in prostate definition with tissue harmonic and brightness mode imaging. Int J Radiat Oncol. https://doi.org/10.1016/j.ijrobp.2011.02.013