Title

Domain generalization for prostate segmentation in transrectal ultrasound images: A multi-center study

Permalink

https://escholarship.org/uc/item/43z0m8s3

Authors

Vesal, Sulaiman

Gayo, Iani

Bhattacharya, Indrani

et al.

Domain Generalization for Prostate Segmentation in Transrectal Ultrasound Images: A Multi-center Study

Sulaiman Vesal[a,*], **Iani Gayo**[b], **Indrani Bhattacharya**[a,c], **Shyam Natarajan**[d], **Leonard S. Marks**[d], **Dean C Barratt**[b], **Richard E. Fan**[a], **Yipeng Hu**[b], **Geoffrey A. Sonn**[a,**], **Mirabela Rusu**[c,**]

[a]Department of Urology, Stanford University, 300 Pasteur Drive, Stanford, CA 94305, USA

[b]Centre for Medical Image Computing, Wellcome/EPSRC Centre for Interventional & Surgical Sciences, and Department of Medical Physics & Biomedical Engineering, University College London, 66-72 Gower St, London WC1E 6EA, UK

[c]Department of Radiology, Stanford University, 300 Pasteur Drive, Stanford, CA 94305, USA

[d]Department of Urology, University of California Los Angeles, 200 Medical Plaza Driveway, Los Angeles, CA 90024, USA

Abstract

Prostate biopsy and image-guided treatment procedures are often performed under the guidance of ultrasound fused with magnetic resonance images (MRI). Accurate image fusion relies on accurate segmentation of the prostate on ultrasound images. Yet, the reduced signal-to-noise ratio and artifacts (e.g., speckle and shadowing) in ultrasound images limit the performance of automated prostate segmentation techniques and generalizing these methods to new image domains is inherently difficult. In this study, we address these challenges by introducing a novel 2.5D deep neural network for prostate segmentation on ultrasound images. Our approach addresses the limitations of transfer learning and finetuning methods (i.e., drop in performance on the original training data when the model weights are updated) by combining a supervised domain adaptation technique and a knowledge distillation loss. The knowledge distillation loss allows the preservation of previously learned knowledge and reduces the performance drop after model finetuning on new datasets. Furthermore, our approach relies on an attention module that considers model feature positioning information to improve the segmentation accuracy. We trained our model on 764 subjects from one institution and finetuned our model using only ten subjects

*Corresponding author. svesal@stanford.edu (Sulaiman Vesal).
**Equally contributed as senior authors.

from subsequent institutions. We analyzed the performance of our method on three large datasets encompassing 2067 subjects from three different institutions. Our method achieved an average Dice Similarity Coefficient (Dice) of 94.0 ± 0.03 and Hausdorff Distance (HD95) of 2.28 *mm* in an independent set of subjects from the first institution. Moreover, our model generalized well in the studies from the other two institutions (Dice: 91.0 ± 0.03; HD95: 3.7 *mm* and Dice: 82.0 ± 0.03; HD95: 7.1 *mm*). We introduced an approach that successfully segmented the prostate on ultrasound images in a multi-center study, suggesting its clinical potential to facilitate the accurate fusion of ultrasound and MRI images to drive biopsy and image-guided treatments.

Keywords

Transrectal Ultrasound; Gland Segmentation; Deep Learning; Prostate MRI; Targeted Biopsy; Continual Learning Segmentation

1. Introduction

Prostate cancer is the third most common cancer diagnosed globally and the fifth leading cause of cancer-related mortality and morbidity in men (Sung et al., 2021). Transrectal ultrasound-guided (TRUS) biopsy procedures are used to diagnose prostate cancer (Michalski et al., 2016; Sarkar and Das, 2016). TRUS images show the prostate in real-time, allowing urologists to guide the biopsy needle during the procedure. Yet, these images have a low signal-to-noise ratio and artifacts (e.g., speckle and shadowing), which reduce the ability of clinicians to reliably distinguish cancerous from normal regions. Therefore, TRUS-guided biopsy procedures usually involve blind sampling of 12 regions throughout the prostate (Harvey et al., 2012). Such blinded systematic biopsy procedures sample <1% of the prostate, missing 52% clinically significant cancer (Williams et al., 2022; Schimmöller et al., 2016).

To address the inability of TRUS images to reliably show cancer, approaches have been developed to fuse magnetic resonance images (MRI) with TRUS images to project suspicious lesions from MRI onto TRUS images and target them during biopsy (T taru et al., 2021; Liau et al., 2019). Fusion requires the registration of MRI and TRUS images which is typically done by aligning the prostate boundary. The prostate boundaries are usually manually outlined on both MRI and TRUS images by clinical experts, usually radiologists for MRI and urologists for TRUS (T taru et al., 2021). Prostate segmentation on TRUS images could also be useful for other applications, including computer-aided diagnosis and targeting cancers using ultrasound images alone.. Accurate manual segmentation of the prostate on TRUS images is a tedious, time-consuming task that suffers from inter- and intra-observer variability due to the reduced quality of the images. Moreover, it is common for the prostate to require segmentation multiple times throughout the biopsy procedure to account for motion and tissue changes (Wang et al., 2019).

Methods for automated prostate segmentation on TRUS images hold the potential to improve accuracy, reduce inter-reader variability and reduce the time required for manual prostate segmentation during clinical procedures. Numerous automated and semi-automated algorithms have been presented for prostate segmentation on TRUS images. Some

approaches used shape statistics as prior knowledge (Li et al., 2016; Ghose et al., 2012), yet required the intervention of an expert user. Other approaches extracted textural features from TRUS images and combined them with traditional machine learning methods to formulate a classification task (Zhan and Shen, 2006; Yang et al., 2016). These methods used hand-crafted features for segmentation, which are inadequate for capturing high-level semantic information and failed to deliver accurate segmentation for complex prostate cases.

Deep learning-based methods have achieved high accuracy in medical image segmentation tasks compared to non-learning methods (Liu et al., 2020; Azizi et al., 2018; Aldoj et al., 2020; Vesal et al., 2021), and have already been used for prostate segmentation on TRUS images (Anas et al., 2017; Ghavami et al., 2018; van Sloun et al., 2021; Jaouen et al., 2019; Girum et al., 2020; Wang et al., 2019; Orlando et al., 2020; Lei et al., 2019; Xu et al., 2021). Most deep learning-based segmentation methods rely on supervised encoder-decoder architectures. Some studies incorporated prior shape information as statistical shape models to improve the segmentation of challenging regions, e.g., apex and base (Zeng et al., 2018; Karimi et al., 2019; Yang et al., 2017). Other studies explored temporal information of TRUS scans using recurrent neural networks (RNNs) (Anas et al., 2018), attention mechanism (Wang et al., 2019) or shadow augmentation (Xu et al., 2021) to improve segmentation quality. However, most studies evaluated their methods on a small set of patients with data from a single institution and a single manufacturer, thus providing limited evidence about generalization across data from other institutions and different imaging devices, vendors, and data acquisition parameters (end-firing and side-firing probes).A recent study (van Sloun et al., 2021) investigated the use of deep learning models for multi-center prostate gland segmentation on TRUS images. The author employed a standard UNet architecture to segment the prostate gland. The model was trained and tested on a small number of patients from three different institutions (a total of 436 TRUS images from 181 men), acquired using only end-fire probes. The reduced training population size and homogeneous data possibly limit the generalizability of their approach, particularly in ultrasound images acquired with a different type of probe or at a different institution.

Deep learning models often require large amounts of data during training to achieve robust and consistent performance across data from multiple institutions. Transfer learning or finetuning (Weiss et al., 2016) is one method for improving the generalization of segmentation models on new data. The main drawback of classical finetuning approach is that it forgets previous knowledge since the model's weights are being updated (Michieli and Zanuttigh, 2019; Cermelli et al., 2020). As a result, the performance deteriorates when the newly trained model is tested on the previous data (Michieli and Zanuttigh, 2019). Knowledge distillation techniques (Wang and Yoon, 2022) have been widely used to preserve the high performance of a model when applied to new tasks. It was originally used to retain the performance of a complex model when adopting to a smaller network for more efficient deployment (Hinton et al., 2015).

Several studies attempted to apply the knowledge distillation technique for a variety of objectives in the computer vision and medical domains, including cross-modality learning (Tian et al., 2020) (transfer knowledge from one modality to another modality without the need of any additional annotations), metric learning (Park et al., 2019; Kim et al., 2021)

(map the input feature representations to an embedding space), and network regularization (Yun et al., 2020) (enhance the generalization performance of deep neural networks using regularization losses). Similarly, approaches based on knowledge distillation have been developed for domain adaptation (Meng et al., 2018; Zhou et al., 2020). The domain adaptation techniques aim to reduce the gap between two domains, which is similar to the domain generalization technique. Moreover, the standard knowledge distillation approaches are based on a teacher-student training scheme, where a teacher model first learns the task and distills the knowledge to a student model, which is more compact with less trainable parameters. Several studies (Meng et al., 2018; Feng et al., 2021; Liang et al., 2022) attempted to train many teachers models in the source domains and ensemble them to distill knowledge into the student approach. However, because these techniques use several pretrained teacher models, they are computationally expensive to train. Recently, the knowledge distillation technique was further adopted for continual learning tasks to keep the network's responses on the previous tasks unchanged while updating it with new training samples (Cermelli et al., 2020). As a result, this helped to reduce the effect of catastrophic forgetting after each round of finetuning on a new task and improve the performance.

In this paper, we present an end-to-end deep learning-based segmentation model for prostate gland segmentation on TRUS images. We introduce an approach for model generalization that utilizes a knowledge distillation loss (Michieli and Zanuttigh, 2019) to mitigate "catastrophic forgetting" while applying the model to images from multiple institutions. We first train a segmentation model on our large cohort of in-house TRUS images (cohort C1, n=764), then finetune the model on subjects from two other institutions (cohorts C2, C3) with relatively few annotated examples by transferring information from the first model to the subsequent models (Fig. 1).

Our study has three main contributions:

- We introduce a deep learning framework for accurate prostate gland segmentation in TRUS images, with the presence of considerable variation in intensity and image acquisition parameters. We improve the generalization capabilities of our model across data from three institutions.

- To limit the effect of catastrophic forgetting during transfer learning, we adapted a training scheme that utilizes knowledge distillation loss during the finetuning process on new data.

- Extensive experiments on multi-center data with different ultrasound probes demonstrate the proposed approach brings substantial gains over existing approaches.

2. Methodology

2.1. Data and cohort description

To develop and validate our prostate gland segmentation algorithm, our study included patients from three different institutions. This retrospective chart review study was approved by the Institutional Review Board (IRB) of Stanford University. As a chart review of

previously collected data, patient consent was waived. The device and clinical characteristics of these data cohorts are listed in Table 1. Examples of TRUS images are shown in Fig. 1.

Cohort C1 included 954 men who underwent TRUS-MRI targeted biopsy utilizing the Artemis biopsy system (Eigen, Grass Valley, CA) at Stanford. Ultrasound scans were carried out using the Hitachi Hi-Vision 5500 7.5 MHz or the Noblus C41V 2-10 MHz end-fire ultrasound probe. The 3D scans were obtained by rotating the end-fire probe 200 degrees around its axis and interpolating to resample the volume to isotropic resolution. The prostate gland was segmented by an expert urologist during the TRUS-MRI targeted biopsy procedures.

Cohort C2 included 1,161 men who underwent biopsy at the University of California Los Angeles (Natarajan et al., 2020; Sonn et al., 2013; Clark et al., 2013). Hitachi Hi-Vision 5500 7.5 MHz end-fire ultrasound probes were used for ultrasound scanning. The volume was resampled with an isotropic resolution by rotating the end-fire probe 200 degrees about its axis and interpolating. This cohort is similar to cohort C1 in terms of the ultrasound image reconstruction method.

Cohort C3: included TRUS scans acquired from 106 men as part of the SmartTarget Biopsy Trial (Ghavami et al., 2018) who underwent targeted transperineal biopsy. For each patient, a continuous rotational 3D acquisition was used to acquire 50-120 sagittal slices to cover the whole prostate. These images were acquired using side-fire ultrasound probes. Three trained biomedical engineering experts segmented the prostate in all images where the prostate gland was visible (Ghavami et al., 2018).

2.2. Data Preprocessing

Multiple preprocessing steps were applied to the TRUS images. Bi-linear interpolation was used to resample the images to the same spacing ($0.25mm \times 0.25mm$) and to resize to 128×160 pixels, while maintaining the aspect ratio. No resampling was performed on the z-axis, and all slices were included in the training. For all studies in the three cohorts, the original TRUS pixel intensities (ranging between [0, 255]) were mapped to the 0-1 range using a min-max normalization. No cropping was applied as our models seek to both localize and segment the prostate. To improve the contrast of the TRUS images, the contrast limited adaptive histogram equalization (Pizer et al., 1990) method was used with a default window size of 4×4 pixels.

Train-test splits: For cohort C1, 764 patients (n=802 TRUS scans) were used for training and validation of the models and 190 patients (n=220 TRUS scans) for testing. Of the 1,151 patients (n=1,761 TRUS scans) in cohort C2, ten patients (n=10 TRUS scans) were randomly chosen for finetuning the model, and the rest of the subjects were used for model testing. Similarly, for cohort C3 (n= 109 TRUS scans), only ten scans were used to finetune the model, as detailed in Table 1. Some patients in cohorts C1 and C2 had multiple TRUS studies.

2.2.1. Segmentation Network Architecture—Leveraging multi-scale information and boundary region guidance can help the segmentation model extract more discriminative

features for robust prostate segmentation, especially given the large variation in the field of view and natural variation in prostate shapes. Thereby, we constructed a 2.5D convolutional network architecture called Coordination Dilated Residual-UNet (CoordDR-UNet) (Fig. 2) inspired by (Vesal et al., 2021). The segmentation model **G** has encoder and decoder paths that are connected by a bottleneck block. Every block in the encoder and decoder paths has two 2D convolution layers followed by a Rectified Linear Unit (ReLU), batch-normalization, and a 2D max-pooling layer to reduce the dimensionality of the feature maps. A residual connection was introduced to each encoder block to optimize the flow of gradients and force the encoder to extract more discriminative features (He et al., 2016). A *softmax* activation function was used in the model's last layer to generate the probability segmentation map of the prostate and background. The bottleneck convolution layers of UNet are replaced by dilated convolutions. This allows the model to collect both global and local contextual information by expanding the receptive field (Wang and Ji, 2018; Vesal et al., 2021). Therefore, we constructed a block of stacked dilated convolutions, the outputs of which are summed. To address the issue of gridding artifacts, each subsequent layer has complete access to earlier features learned using different dilation rates. In our model configuration, we employed four dilated convolutions in the model bottleneck with a dilation rate of $1 - 8$.

Moreover, our approach relies on attention mechanisms to force the deep learning model to pay more attention to the uncertain regions during model back-propagation(Schlemper et al., 2019; Hu et al., 2018; Roy et al., 2019), specifically caused by the absence of clear boundaries between the prostate and the surrounding tissue. We attached a coordinate attention block (CAB) (Hou et al., 2021) to our 2.5D DR-UNet to assist the model in improving the expressive power of the learned features. The coordinate attention block takes the output of each encoder block as the input $X_\theta = [x_1, x_2, ..., x_N] \in \mathbb{R}^{h \times w \times n}$ and outputs a transformed tensor with augmented representations, i.e., feature maps, $\mathcal{Y}_\theta = [y_1, y_2, ..., y_N]$. Here n denotes the number of feature maps for each encoder block. There are two 1D global average pooling layers of x_θ and y_θ to encode each channel along with the horizontal and vertical coordinates. Each average pooling concentrates on one coordinate direction and combines information from two spatial directions, providing a pair of feature maps that are direction-aware (Hu et al., 2018).

2.3. Continual Prostate Gland Segmentation

In supervised domain adaptation or finetuning for semantic segmentation, we are given a set of source TRUS images (e.g., from Cohort C1) and their corresponding mask labels in the source domain $\mathbf{D}_s = (\mathbf{I}_s^i, \mathbf{Y}_s^i)_{i=1}^{m_s}$, where $\mathbf{I}_s \in \mathbb{R}^{w \times h \times 3}$ is the stack of three consecutive slices in the TRUS exam, $\mathbf{Y}_s \in \mathbb{R}^{w \times h \times c}$ is the prostate segmentation corresponding to $\mathbf{I}_s$, m_s is the number of source images and $c = 2$ is the number of labels, i.e., prostate and background. In the target site, we are given few labelled images $\mathbf{D}_t = (\mathbf{I}_t^i)_{i=1}^{m_t}$, where m_t is the number of target images. Our goal is to train a supervised model on $\mathbf{D}_s$ and transfer information from $\mathbf{D}_s$ to reduce the gap between two domains, and improve segmentation the accuracy on $\mathbf{D}_t$.

We were motivated by the work of (Michieli and Zanuttigh, 2019) to develop a pipeline that not only produced good segmentation accuracy for the prostate gland but also reduced

the impact of model weight changes during the supervised domain adaptation process. Our approach has three steps.

- First, we trained in a supervised fashion the segmentation model $\mathbf{M}_s$ that segments the prostate in TRUS images using the training data from cohort C1, also referred here as $\mathbf{D}_s$ (Fig. 3). The input to the model $\mathbf{M}_s$ is $\mathbf{I}_s$, which includes a TRUS slice and two neighboring slices from the 3D TRUS volume to construct a three-channel input. The model is trained in a supervised fashion with a multi-class loss function. After training, we saved the obtained model weights as $\mathbf{M}_s$.

- Second, we trained a new model $\mathbf{M}_{t1}$ based on $\mathbf{M}_s$ that takes as the input the TRUS images from cohort C2 (10 annotated cases) for finetuning. Here, the input images were fed to both models $\mathbf{M}_{t1}$ and $\mathbf{M}_s$ while the weights for $\mathbf{M}_s$ were frozen. Our goal was to distill knowledge from the learned model $\mathbf{M}_s$ to $\mathbf{M}_{t2}$ by enforcing the consistency of the latent feature space. This was achieved by minimizing the distance between z and z' using knowledge distillation loss function as a regularization term along with a supervised segmentation loss function.

- For the third step, we repeated step two by creating a new model $\mathbf{M}_{t2}$ based on trained model $\mathbf{M}_{t1}$, which takes as the input the TRUS images from cohort C3 (10 annotated cases). In this step, we distill knowledge from learned model $\mathbf{M}_{t1}$ to $\mathbf{M}_{t2}$ similar to step two by minimizing the distance between the latent feature space.

2.3.1. Loss functions

Segmentation Loss: The segmentation model is trained with a soft-Dice loss. To segment a TRUS image $\mathbf{I}_s \in \mathbb{R}^{h \times w \times 3}$ the output of *Softmax* layer is two probability maps for classes $k = 0, 1$ (background and prostate) where for each pixel $\Sigma_c \mathbf{Y}_{n,k} = 1$. Given the ground-truth label $\mathbf{Y}_{n,k}$ for that identical pixel, the soft Dice loss is computed as follows:

$$\mathscr{L}_{seg}(\mathbf{Y}, \widehat{\mathbf{Y}}) = 1 - \frac{1}{K}\left(\sum_k \frac{2\sum_n \mathbf{Y}_{nk}\widehat{\mathbf{Y}}_{nk}}{\sum_n \mathbf{Y}_{nk} + \sum_n \widehat{\mathbf{Y}}_{nk}} \right) \tag{1}$$

Feature-Space Knowledge Distillation Loss: Knowledge distillation (Hinton et al., 2015) is an efficient technique for transferring knowledge from a well-trained model to a model with limited annotated data. It also has been widely used in continual learning frameworks (Michieli and Zanuttigh, 2019) when a trained model is updated on a new task that interferes with the learned representations on the previous task. We hypothesize that the same catastrophic forgetting problem exists for domain generalization across multi-institutional data. The $\mathbf{M}_s$ model (global model) is updated on a small set of randomly sampled data for each cohort at each round, while the cohorts have different distributions from the previous round. Therefore, to reduce the impact of model weight changes in the latent feature space of $\mathbf{M}_{t1}$ and $\mathbf{M}_{t2}$ during finetuning, an L2-norm was applied as a knowledge distillation loss.

This regularization loss enforced the model to preserve the previous knowledge and keep the latent features space of $\mathbf{M}_{t_n}$ the same, where n is the number of available cohorts.

$$\mathscr{L}_{KDL} = \frac{\|\mathbf{E}_1(z) - \mathbf{E}_2(z')\|_2^2}{|\mathbf{D}_t|} \tag{2}$$

where z and z' denotes the latent features computed by model encoder $\mathbf{E}_1$ and $\mathbf{E}_2$ when a TRUS image from cohort C2 or C3 is fed.

The overall loss function for training the whole pipeline is defined as:

$$\mathscr{L}_S = L_{seg} + \lambda L_{KDL} \tag{3}$$

where λ is the weight that controls the impact of knowledge distillation term during optimization.

2.4. Implementation details

The prostate gland segmentation framework was trained in an end-to-end fashion using the Adam optimizer with an initial learning rate of $\eta = 10^{-3}$ and exponential weight decay $\alpha = 0.01$ for 500 epochs, with a batch size of 64. The momentum parameters for the Adam optimizer were set to $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The learning rate was reduced on the plateau if the validation loss did not improve after every 30 epochs. We used early stopping if no reduction in validation loss was seen for 50 epochs. During training, we utilized an online data augmentation library (imgAug Jung et al. (2020)) and employed a set of random data augmentation for each input image $\mathbf{I}_s \in \mathbb{R}^{h \times w \times 3}$. The augmentation includes horizontal/vertical axis flipping, image scaling with a factor of [0.8, 1.2], Gaussian noise with $\sigma \in [0.0, 0.3]$, elastic deformation with the displacement field strength of $\alpha \in [0.5, 0.35]$ and a displacement field smoothness of $\sigma = 0.25$. To evaluate our method, we compared the resulting segmentations with the ground-truth masks using several metrics namely Dice similarity coefficient (Dice), Hausdorff Distances (HD), and sensitivity. We used the 95th percentile of the Hausdorff distances (HD95) between model prediction $\mathbf{Y}$ and ground-truth mask $\widehat{\mathbf{Y}}$ to eliminate the impact of outliers. The paired Student's t-test was used to assess the statistical significance of Dice differences when comparing multiple methods (Nadeau and Bengio, 2003).

We used PyTorch 1.7 (Paszke et al., 2019) to develop and train all models. Training took around 24 hours on an NVIDIA RTX A6000 GPU with 32GB memory. The code is available online at: https://github.com/pimed/TRUSGlandSegmentation

3. Results

Several experiments were carried out to 1) determine the best segmentation model on cohort C1-test and 2) identify the best strategy for knowledge transfer during finetuning on data from other institutions.

3.1. Prostate Gland Segmentation

We compared CoordDR-UNet with several well-known segmentation approaches, including UNet, Attention-UNet, Nested-UNet, Dilated-residual UNet, and Deep Attentive Features Network (DAFNet) (Wang et al., 2019) which is a new approach for 3D segmentation of TRUS. All five approaches were trained using their public implementations, and their training parameters were adjusted to obtain the best segmentation results. Fig. 4 shows four slices of a test set patient in cohort C1 from apex to base (column (a)). It also shows comparative performance among the UNet (column (b)), Attention-UNet (column (c)), Nested-UNet (column (d)), DAFNet (column (e)) and CoordDR-UNet (column (f)). CoordDR-UNet segmented the prostate successfully, with better prediction at the apex and the base of the prostate (rows 1,4). Furthermore, for midgland slices (row 2-3), CoordDR-UNet produced very accurate prostate gland mask predictions compared to other state-of-the-art methods.

Table 2 summarizes the results for cohort C1-test for all models using several segmentation metrics. To speed up volumetric segmentation while minimizing memory requirements and incorporating contextual and temporal information for the model, all models were trained with a 2.5D input. As seen in Table 2, the 2.5D UNet outperformed the 2D and 3D UNet with a Dice score of 0.91 ± 0.07. The surface distance metric, HD95, was also reduced, i.e., 3.0 *mm* for 2.5D UNet vs 4.03 *mm* for 2D UNet. The performance of Attention-UNet and Nested-UNet was similar to that of UNet. DAFNet, which is a recent model specifically designed for prostate gland segmentation in 3D TRUS achieved a Dice score of 0.92 ± 0.03 and HD95 value of 2.87 *mm.* Our proposed CoordDR-UNet approach, which incorporates coordinate attention blocks and dilated convolutions, outperformed other models, with a Dice score of 0.94 ± 0.03 and the lowest HD95 score of 2.29 *mm* compared with the other approaches. We will refer to this model as the model $\mathbf{M}_s$. Figure S1 in the supplementary material shows the impact of training data size with respect to segmentation accuracy.

To show the benefit of the 2.5D input representations, we also trained the CoordDR-UNet model with 2D and 3D data representations as input. For the 3D model, the standard architecture of CoordDR-UNet was changed by replacing all 2D operations with 3D operations, including convolution, batch normalization, max-pooling, and upsampling layers. Moreover, the 3D TRUS images were downsampled to a fixed size ($80 \times 160 \times 128$) similar to (Wang et al., 2019). The quantitative results (Table 2) show that CoordDR-UNet 2.5D outperformed the models trained with 2D and 3D input representations, suggesting the benefit of the 2.5D representation. CoordDR-UNet with 2D input achieved a Dice score of 0.92 ± 0.03 and HD95 value of 3.29 *mm* while CoordDR-UNet with 3D input obtained a Dice score of 0.92 ± 0.04 and HD95 value of 3.00 *mm.* Furthermore, Fig. 5 shows visual comparison between CoordDR-UNet models trained with 2D, 2.5D and 3D input representations. One limitation of the 2D segmentation approaches for prostate gland segmentation is inconsistency across adjacent slices. The difference between CoordDR-UNet 2.5D and CoordDR-UNet with 2D and 3D input representations is shown in Fig. 6. The 2.5D representation achieved a more accurate segmentation and reduced inconsistencies across adjacent slices. The 3D segmentation approach has smoother boundaries, but the output is less accurate when compared to the ground-truth masks. These results highlight

the benefit of 2.5D representations, which not only produced highly accurate prostate gland segmentation but also consistent segmentation across adjacent slices.

For the error analysis, we computed the surface error between model predictions and their corresponding ground truth. Fig. 7 shows the 3D visualization of the surface distance between the ground-truth and segmentation output by different methods for two test cases from the cohort C1-test. Our method consistently achieved accurate and robust segmentation covering the whole prostate, including challenging regions such as the apex and base. CoordDR-UNet 2.5D had the lowest average HD95 surface distance (2.29 mm) compared to other approaches for cohort C1-test data. DAFNet produced a more smooth segmentation output because it was trained with 3D input representations. However, it had a higher surface distance error (2.87 mm) because the segmentation output is less accurate in comparison to ground truth.

3.2. Model performance on independent data

To assess model generalizability, we evaluated our framework on TRUS images from cohorts C2 and C3 (Table 3). We considered four different scenarios to demonstrate the performance of CoordDR-UNet 2.5D with knowledge distillation. In the first scenario, we directly tested the model M_1 (trained with data from Cohort C1) on cohorts C2 and C3 without domain adaptation or finetuning. The model achieved good results on cohort C2 (Table 3, 3rd row) with a Dice score of 0.89 ± 0.03 (P 0.05) and HD95 of 4.03 mm. This relatively high performance is due to the similar data acquisition and vendor for the TRUS images in cohorts C1 and C2. However, the performance of our model M_1 on cohort C3 was poor (Dice score = 0.24 ± 0.29 (P 0.05)). This reduced performance is likely caused by the differences in acquisition, field of view, ultrasound manufacturer, and ultrasound probes (side-fire) between cohorts C1 and C2 versus C3 (Fig. 10).

In the second scenario, we obtained ultrasound scans and ground truth from the prostate segmentation from ten random subjects in each cohort (C2 and C3). The CoordDR-UNet model was trained from scratch (without pretraining) to investigate how the model performs with limited annotated data. On cohort C2-test data, this model (Table 3, 4th row) obtained a Dice score of 0.71 ± 0.09 and HD95 value of 14.8 mm, which demonstrates a substantial performance drop in comparison to the CoordDR-UNet trained on cohort C1 data. This is because a model with only ten training samples cannot capture all of the variations in the test data since prostate volumes and image quality vary significantly between patients. We observed similar results when we utilized only ten examples from cohort C3 to train CoordDR-UNet from scratch. Furthermore, the model without pretraining (Table 3, 8th row) achieved a Dice score of 0.78 ± 0.14 (P 0.05) and surface distance of 10.23 mm (HD95) on cohort C3-test. Training with only ten TRUS scans from cohort C3 assisted the model to somewhat capture the overall prostate shape and geometry compared to only using the model M_s which had not seen any data from Cohort C3.

In the third scenario, we performed standard finetuning on CoordDR-UNet with a pretrained model and without distillation loss. All the layers in CoordDR-UNet were frozen except the last layer (Weiss et al., 2016). On cohort C2-test data, this model obtained a Dice score of 0.90 ± 0.03 (P 0.05) and HD95 value of 3.80 mm (Table 3, 5th row). On cohort C3-test

data, this model achieved a Dice score of 0.80 ± 0.03 (P 0.05) and HD95 value of 7.30 *mm* (Table 3, 10th row).

In the fourth scenario, we tested CoordDR-UNet with a pretrained model and knowledge distillation loss (Table 3). In the studies from cohort C2-test, the model obtained a Dice score of 0.91 ± 0.03 and HD95 value of 3.69 *mm* (Table 3, 6th row), showing similar results with studies on cohort C1-test data. We refer to this model as model $\mathbf{M}_{t1}$. Moreover, CoordDR-UNet + KDL improved the prostate gland segmentation for studies in cohort C3-test, achieving a Dice score of 0.82 ± 0.16 (P 0.05), and HD95 value of 7.13 *mm*. We refer to this model as model $\mathbf{M}_{t2}$. The quantitative evaluation showed that the CoordDR-UNet + KDL approach achieved a significantly higher (P 0.05) Dice score compared to the CoordDR-UNet without pretraining and the CoordDR-UNet direct prediction for both cohorts (Fig. 8).

Figs. 9-10 display segmentation outputs for the best, average, and worst performing cases in cohorts C2-test and C3-test using the different generalization approaches we considered. In data from both cohorts, CoordDR-UNet + KDL obtained the most accurate segmentation among different training strategies. Fig. 11 shows 3D segmentation results for CoordDR-UNet on TRUS images from the three cohorts, as well as the corresponding surface distance between segmented surfaces and ground truth volumes. These results highlight that the proposed method obtained accurate and smooth segmentation surfaces covering the whole prostate region for cohorts C1-C2 (columns a-b), but the surface distances are high for the test cases in cohort C3 (column c).

3.3. Model performance after finetuning

To evaluate the performance of our method after all iterations of finetuning on cohorts C1, C2, and C3. Table S1 in the Supplementary material shows the quantitative results for the following three models:

- $\mathbf{M}_s$: model trained using the training data from cohort C1

- $\mathbf{M}_{t1}$: model $\mathbf{M}_s$, finetuned using 10 TRUS images from cohort C2 using CoorDR-UNet and Knowledge Distillation Loss

- $\mathbf{M}_{t2}$: model $\mathbf{M}_{t1}$, finetuned using 10 TRUS images from cohort C3 using CoorDR-UNet and Knowledge Distillation Loss

When training $\mathbf{M}_{t1}$ by finetuning $\mathbf{M}_s$ with knowledge distillation loss in the 10 cases from cohort C2, we tested the model $\mathbf{M}_{t1}$ on the studies in cohort C2-test and C3-test, but as well in studies in cohort C1-test to assess whether the model suffered from catastrophic forgetting. Based on the results, the final model ($\mathbf{M}_{t2}$), even though had a small drop of performance on the test studies in cohorts C1 and C2, achieved the overall best performance across all three data cohorts with a Dice score of 0.91, 0.88 and 0.82 for the test studies in cohorts C1, C2 and C3 (Fig. 12).

3.4. Ablation Study

Impact of Input Data: We conducted an ablation study and trained the CoordDR-UNet 2.5D with a different number of slices $c \in [1, 3, 5, 7]$ as the input to determine the optimal number of neighboring slices. Fig. 13a shows the Dice score on cohort C1-test data for the different values of c. The model trained with a single (2D) slice as input and no neighboring slices obtained a Dice score of 0.92. However, with three slices as input data, the CoordDR-UNet 2.5 achieved a Dice score of 0.94. Furthermore, increasing the number of neighboring slices in the CoordDR-UNet 2.5 model did not improve the segmentation accuracy and slightly decreased the Dice score. Therefore, we trained all 2.5D models with three slices as input data. Also, previous work on prostate gland segmentation in MRI (Soerensen et al., 2021) showed that having a 2.5D model with three adjacent slices is successful in generating accurate segmentation on prostate MRI.

Impact of Finetuning Cases: To evaluate the impact of training data size for model finetuning and its influence on domain generalization across cohort C2 and C3 test data, we conducted experiments by finetuning CoordDR-UNet+KDL with the various number of cases $I \in [0, 2, 4, 6, 8, 10, 12, 14]$. Fig. 13b shows the segmentation Dice coefficient score of CoordDR-UNet+KDL for cohort C2-test and C3-test. The model fine-tuned with two subjects achieved a Dice score of 0.90, and adding more cases, only slightly improved the segmentation performance (increase to 0.91). However, finetuning with 10, 12, and 14 subjects have a similar Dice score of 0.91. These results imply that when TRUS images from a new institution are acquired using a similar ultrasound vendor and probe to training data, just a few instances are required to achieve satisfactory prostate gland segmentation. For cohort C3-test data, finetuning with more data showed a significant improvement in Dice score. Since the TRUS images in this cohort are significantly different and acquired using a side-fire probe. The model without any finetuning subject achieved a Dice score of 0.43, however, with adding more cases, the Dice score increased. For instance, the model finetuned with 12 subjects achieved a Dice score of 0.84.

Impact of Lambda: To control the influence of knowledge distillation loss during finetuning, we used λ as a weight factor. If we set $\lambda = 0$, then the model is trained using the Classical finetuning scenario with no knowledge distillation. To find out the impact of λ, we finetuned CoordDR-UNet+KDL model on cohort C2 training data (10 cases) with various λ values. Fig. 13c shows the line-plot and the computed Dice value for a range of λ values [0 – 1.0]. The value of $\lambda = 0.2$ achieved the best Dice value on the test data and increasing the λ resulted in a reduction in segmentation accuracy. Therefore, we set the $\lambda = 0.2$ for the rest of the experiments.

4. Discussion

In this study, we introduced a deep learning model for prostate gland segmentation in 3D TRUS scans called Coordination Dilated-Residual UNet (CoordDR-UNet), as well as a strategy for domain generalization which was tested using TRUS images from three different institutions. Our experiments demonstrated that the proposed method accurately localized and segmented the prostate gland in the presence of variations in image acquisition

parameters and generalized well on data from multiple institutions. These encouraging results are attributable to the addition of a coordination attention block in our model that increased segmentation accuracy in ambiguous regions (e.g., prostate borders at the apex). Moreover, we introduced a model generalization approach for prostate gland segmentation in 3D TRUS based on distilling knowledge from previously trained models. The primary goal of using the knowledge distillation strategy during finetuning was to reduce the impact of catastrophic forgetting, which manifests as a performance loss after each finetuning cycle on data from a new center. The experiment results showed (Table 3) that when the CoordDR-UNet was finetuned on cohort C3 using the Classical finetuning approach, it obtained a Dice score of 0.80 ± 0.03 (P 0.05) and HD95 value of 7.30 *mm*, whereas the CoordDR-UNet+KDL approach increased the Dice score by 2.0% (0.82 ± 0.16. p 0.05). Similarly, for cohort C2, the CoordDR-UNet+KDL outperformed the Classical finetuning approach.

Furthermore, we compared our method with existing segmentation methods. While a direct comparison with published results from prior studies was not feasible due to a lack of data availability, we trained the prior algorithms on our cohort and designed fair comparative experiments. Our proposed model outperformed other models in several evaluation metrics (e.g., the boundary distance between the prediction and ground-truth for all cases in the test cohorts was on average only 4.37 *mm* versus 7.9 *mm* for alternative approaches).

Our proposed generalization framework has both similarities and differences to federated learning approaches (Kairouz et al., 2021; Liu et al., 2021). Similar to federated learning, we transferred our learned model weights and training code to an institution without directly accessing their data. The model was then trained, finetuned, and tested locally, with the results being reported back. Yet, unlike federated learning, we did not apply privacy-preserving techniques, and the weights aggregation was solely based on retraining with new data. Further work will involve testing the model on data from additional institutions without data sharing.

Our study has four limitations. First, we only segmented the whole prostate gland without segmenting the distinct anatomic zones, which have the potential for other tasks beyond TRUS-MRI fusion. Second, our study includes images from only two ultrasound vendors. Future studies will expand this work to include TRUS scans from other ultrasound vendors. Third, while our proposed method achieved encouraging results on the studies in cohort C3-test, there is still potential for further improvement. Adopting additional data augmentation strategies might assist in addressing issues related to variations in field of view and ultrasound probes such as side-fire and end-fire. Fourth, the number of samples chosen from C2 and C3 (10 patients) to finetune the models can be ablated to find the optimal number of patients for better segmentation accuracy.

Our proposed approach for automated prostate segmentation on TRUS images can improve clinical workflow in four ways. First, our approach aids in the diagnosis of prostate cancer by enabling a better biopsy procedure with more accurately registered MRI-TRUS images and therefore possibly better needle targeting during biopsy (the registrations are driven by the segmentation of the prostate provided by our approach). The MRI-TRUS fusion

step during the targeted biopsy is prone to errors (Das et al., 2020; Avola et al., 2021), and registration accuracy is affected by the quality of prostate gland segmentation on both TRUS and MR images. Second, our generalizable approach is capable of producing high-quality segmentation for a wide range of probes, and it can be integrated into TRUS scanners as an alternative method for prostate segmentation. Third, our approach provides objective, reproducible, and fast estimates of prostate volume (run-time including pre- and post-processing steps: 12 seconds), which has been reported to be operator-dependent, difficult to replicate, and less accurate (van Sloun et al., 2021). Fourth, our approach allows for better planning of focal treatment to mark the prostate capsule while sparing the tissue beyond the prostate.

5. Conclusion

We have introduced an accurate and generalizable approach for prostate segmentation in 3D transrectal ultrasound images to limit the need for manual segmentation of the prostate during targeted biopsy procedures. The proposed deep learning framework outperformed state-of-the-art segmentation approaches in accuracy and generalization and was tested on data from three institutions. In comparison to traditional approaches that require substantial user input and processing time, our pipeline delivered accurate and efficient segmentation of the prostate without any user input. The ease of use and speed of our pipeline make it appealing for practical deployment to allow direct segmentation of the prostate during biopsy or treatment procedures. In the future, we would like to evaluate our method on other tasks and imaging modalities, e.g., prostate segmentation in MRI or CT.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

References

Aldoj N, Biavati F, Michallek F, Stober S, Dewey M, 2020. Automatic prostate and prostate zones segmentation of magnetic resonance images using densenet-like u-net. Scientific Reports 10, 14315. [PubMed: 32868836]

Anas EMA, Mousavi P, Abolmaesumi P, 2018. A deep learning approach for real time prostate segmentation in freehand ultrasound guided biopsy. Medical Image Analysis 48, 107–116. [PubMed: 29886268]

Anas EMA, Nouranian S, Mahdavi SS, Spadinger I, Morris WJ, Salcudean SE, Mousavi P, Abolmaesumi P, 2017. clinical target-volume delineation in prostate brachytherapy using residual neural networks, in: Medical Image Computing and Computer Assisted Intervention, MICCAI 2017, pp. 365–373.

Avola D, Cinque L, Fagioli A, Foresti G, Mecca A, 2021. Ultrasound medical imaging techniques: A survey. ACM Comput. Surv 54.

Azizi S, Van Woudenberg N, Sojoudi S, Li M, Xu S, Abu Anas EM, Yan P, Tahmasebi A, Kwak JT, Turkbey B, Choyke P, Pinto P, Wood B, Mousavi P, Abolmaesumi P, 2018. Toward a real-time system for temporal enhanced ultrasound-guided prostate biopsy. International Journal of Computer Assisted Radiology and Surgery 13, 1201–1209. [PubMed: 29589258]

Cermelli F, Mancini M, Rota Bulò S, Ricci E, Caputo B, 2020. Modeling the background for incremental learning in semantic segmentation, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9230–9239.

Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, Moore S, Phillips S, Maffitt D, Pringle M, Tarbox L, Prior F, 2013. The cancer imaging archive (tcia): Maintaining and operating a public information repository. Journal of Digital Imaging 26, 1045–1057. [PubMed: 23884657]

Das CJ, Razik A, Netaji A, Verma S, 2020. Prostate mri–trus fusion biopsy: a review of the state of the art procedure. Abdominal Radiology 45, 2176–2183. [PubMed: 31897683]

Feng H, You Z, Chen M, Zhang T, Zhu M, Wu F, Wu C, Chen W, 2021. Kd3a: Unsupervised multi-source decentralized domain adaptation via knowledge distillation, in: Meila M, Zhang T (Eds.), Proceedings of the 38th International Conference on Machine Learning, PMLR. pp. 3274–3283.

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. Journal of Medical Imaging 6, 1–6.

Ghose S, Oliver A, Martí R, Lladó X, Vilanova JC, Freixenet J, Mitra J, Sidibé D, Meriaudeau F, 2012. A survey of prostate segmentation methodologies in ultrasound, magnetic resonance and computed tomography images. Computer Methods and Programs in Biomedicine 108, 262–287. [PubMed: 22739209]

Girum KB, Lalande A, Hussain R, Créhange G, 2020. A deep learning method for real-time intraoperative us image segmentation in prostate brachytherapy. International Journal of Computer Assisted Radiology and Surgery 15, 1467–1476. [PubMed: 32691302]

Harvey CJ, Pilcher J, Richenberg J, Patel U, Frauscher F, 2012. Applications of transrectal ultrasound in prostate cancer. The British Journal of Radiology 85, S3–S17. [PubMed: 22844031]

He K, Zhang X, Ren S, Sun J, 2016. Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778.

Hinton G, Vinyals O, Dean J, 2015. Distilling the knowledge in a neural network, in: NIPS Deep Learning and Representation Learning Workshop.

Hou Q, Zhou D, Feng J, 2021. Coordinate attention for efficient mobile network design, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 13713–13722.

Hu J, Shen L, Sun G, 2018. Squeeze-and-excitation networks, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7132–7141.

Jaouen V, Bert J, Mountris KA, Boussion N, Schick U, Pradier O, Valeri A, Visvikis D, 2019. Prostate volume segmentation in trus using hybrid edge-bhattacharyya active surfaces. IEEE Transactions on Biomedical Engineering 66, 920–933. [PubMed: 30113888]

Jung AB, Wada K, Crall J, Tanaka S, Graving J, Reinders C, Yadav S, Banerjee J, Vecsei G, Kraft A, Rui Z, Borovec J, Vallentin C, Zhydenko S, Pfeiffer K, Cook B, Fernández I, De Rainville FM, Weng CH, Ayala-Acevedo A, Meudec R, Laporte M, et al., 2020. imgaug. https://github.com/aleju/imgaug. Online; accessed 01-Sep-2021.

Kairouz P, McMahan HB, Avent B, Bellet A, Bennis M, et al. , A.N.B., 2021. Advances and open problems in federated learning. Foundations and Trends® in Machine Learning 14, 1–210.

Karimi D, Zeng Q, Mathur P, Avinash A, Mahdavi S, Spadinger I, Abolmaesumi P, Salcudean SE, 2019. Accurate and robust deep learning-based segmentation of the prostate clinical target volume in ultrasound images. Medical Image Analysis 57, 186–196. [PubMed: 31325722]

Kim D, Yoo Y, Park S, Kim J, Lee J, 2021. Selfreg: Self-supervised contrastive regularization for domain generalization, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9619–9628.

Lei Y, Tian S, He X, Wang T, Wang B, Patel P, Jani AB, Mao H, Curran WJ, Liu T, Yang X, 2019. Ultrasound prostate segmentation based on multidirectional deeply supervised v-net. Medical Physics 46, 3194–3206. [PubMed: 31074513]

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. Medical Physics 43, 3090–3103. [PubMed: 27277056]

Liang X, Wu L, Li J, Qin T, Zhang M, Liu TY, 2022. Multi-teacher distillation with single model for neural machine translation. IEEE/ACM Transactions on Audio, Speech, and Language Processing 30, 992–1002. doi:10.1109/TASLP.2022.3153264.

Liau J, Goldberg D, Arif-Tiwari H, 2019. Prostate cancer detection and diagnosis: Role of ultrasound with mri correlates. Current Radiology Reports 7, 7.

Liu Q, Chen C, Qin J, Dou Q, Heng PA, 2021. Feddg: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1013–1023.

Liu Q, Dou Q, Yu L, Heng PA, 2020. Ms-net: Multi-site network for improving prostate segmentation with heterogeneous mri data. IEEE Transactions on Medical Imaging 39, 2713–2724. [PubMed: 32078543]

Meng Z, Li J, Gong Y, Juang BH, 2018. Adversarial teacher-student learning for unsupervised domain adaptation, in: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5949–5953. doi:10.1109/ICASSP.2018.8461682.

Michalski JM, Pisansky TM, Lawton CA, Potters L, 2016. Chapter 53 - prostate cancer, in: Clinical Radiation Oncology (Fourth Edition). fourth edition ed.. Elsevier, Philadelphia, pp. 1038–1095.e18.

Michieli U, Zanuttigh P, 2019. Incremental learning techniques for semantic segmentation, in: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), pp. 3205–3212.

Nadeau C, Bengio Y, 2003. Inference for the generalization error. Machine Learning 52, 239–281.

Natarajan S, Priester A, Margolis D, Huang J, Marks L, 2020. Prostate MRI and Ultrasound With Pathology and Coordinates of Tracked Biopsy (Prostate-MRI-US-Biopsy). doi:10.7937/TCIA.2020.A61IOC1A.

Orlando N, Gillies DJ, Gyacskov I, Fenster A, 2020. Deep learning-based automatic prostate segmentation in 3D transrectal ultrasound images from multiple acquisition geometries and systems, in: Medical Imaging 2020: Image-Guided Procedures, Robotic Interventions, and Modeling, International Society for Optics and Photonics. SPIE. pp. 651–656.

Park W, Kim D, Lu Y, Cho M, 2019. Relational knowledge distillation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3967–3976.

Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, Desmaison A, Kopf A, Yang E, De-Vito Z, Raison M, Tejani A, Chilamkurthy S, Steiner B, Fang L, Bai J, Chintala S, 2019. Pytorch: An imperative style, high-performance deep learning library, in: Advances in Neural Information Processing Systems 32. Curran Associates, Inc., pp. 8024–8035.

Pizer S, Johnston R, Ericksen J, Yankaskas B, Muller K, 1990. Contrast-limited adaptive histogram equalization: speed and effectiveness, in: [1990] Proceedings of the First Conference on Visualization in Biomedical Computing, pp. 337–345. doi:10.1109/VBC.1990.109340.

Ronneberger O, Fischer P, Brox T, 2015. U-net: Convolutional networks for biomedical image segmentation, in: Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015, pp. 234–241.

Roy AG, Navab N, Wachinger C, 2019. Recalibrating fully convolutional networks with spatial and channel "squeeze and excitation" blocks. IEEE Transactions on Medical Imaging 38, 540–549. [PubMed: 30716024]

Sarkar S, Das S, 2016. A review of imaging methods for prostate cancer detection: Supplementary issue: Image and video acquisition and processing for clinical applications. Biomedical Engineering and Computational Biology 7s1, BECB.S34255.

Schimmöller L, Blondin D, Arsov C, Rabenalt R, Albers P, Antoch G, Quentin M, 2016. Mri-guided in-bore biopsy: Differences between prostate cancer detection and localization in primary and secondary biopsy settings. American Journal of Roentgenology 206, 92–99. [PubMed: 26700339]

Schlemper J, Oktay O, Schaap M, Heinrich M, Kainz B, Glocker B, Rueckert D, 2019. Attention gated networks: Learning to leverage salient regions in medical images. Medical Image Analysis 53, 197–207. [PubMed: 30802813]

van Sloun RJ, Wildeboer RR, Mannaerts CK, Postema AW, Gayet M, Beerlage HP, Salomon G, Wijkstra H, Mischi M, 2021. Deep learning for real-time, automatic, and scanner-adapted prostate (zone) segmentation of transrectal ultrasound, for example, magnetic resonance imaging; transrectal ultrasound fusion prostate biopsy. European Urology Focus 7, 78–85. [PubMed: 31028016]

Soerensen SJC, Fan RE, Seetharaman A, Chen L, Shao W, Bhattacharya I, hun Kim Y, Sood R, Borre M, Chung BI, To'o KJ, Rusu M, Sonn GA, 2021. Deep learning improves speed and accuracy of prostate gland segmentations on magnetic resonance imaging for targeted biopsy. Journal of Urology 0, 10.1097/JU.0000000000001783.

Sonn GA, Natarajan S, Margolis DJ, MacAiran M, Lieu P, Huang J, Dorey FJ, Marks LS, 2013. Targeted biopsy in the detection of prostate cancer using an office based magnetic resonance ultrasound fusion device. The Journal of Urology 189, 86–92. [PubMed: 23158413]

Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F, 2021. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for clinicians 71, 209–249. [PubMed: 33538338]

Tian Y, Krishnan D, Isola P, 2020. Contrastive representation distillation, in: International Conference on Learning Representations.

T taru OS, Vartolomei MD, Rassweiler JJ, Virgil O, Lucarelli G, Porpiglia F, Amparore D, Manfredi M, Carrieri G, Falagario U, Terracciano D, de Cobelli O, Busetto GM, Giudice FD, Ferro M, 2021. Artificial intelligence and machine learning in prostate cancer patient management—current trends and future perspectives. Diagnostics 11.

Vesal S, Gu M, Maier A, Ravikumar N, 2021. Spatio-temporal multi-task learning for cardiac mri left ventricle quantification. IEEE Journal of Biomedical and Health Informatics 25, 2698–2709. [PubMed: 33351771]

Wang L, Yoon KJ, 2022. Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks. IEEE Transactions on Pattern Analysis and Machine Intelligence 44, 3048–3068. doi:10.1109/TPAMI.2021.3055564. [PubMed: 33513099]

Wang Y, Dou H, Hu X, Zhu L, Yang X, Xu M, Qin J, Heng PA, Wang T, Ni D, 2019. Deep attentive features for prostate segmentation in 3d transrectal ultrasound. IEEE Transactions on Medical Imaging 38, 2768–2778. [PubMed: 31021793]

Wang Z, Ji S, 2018. Smoothed dilated convolutions for improved dense prediction, in: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, ACM. pp. 2486–2495.

Weiss K, Khoshgoftaar TM, Wang D, 2016. A survey of transfer learning. Journal of Big Data 3, 9.

Williams C, Ahdoot M, Daneshvar MA, Hague C, Wilbur AR, Gomella PT, Shih J, Khondakar N, Yerram N, Mehralivand S, Gurram S, Siddiqui M, Pinsky P, Parnes H, Merino M, Wood B, Turkbey B, Pinto PA, 2022. Why does magnetic resonance imaging-targeted biopsy miss clinically significant cancer? Journal of Urology 207, 95–107. [PubMed: 34433302]

Xu X, Sanford T, Turkbey B, Xu S, Wood BJ, Yan P, 2021. Shadow-consistent semi-supervised learning for prostate ultrasound segmentation. IEEE Transactions on Medical Imaging , 1–1.

Yang X, Rossi PJ, Jani AB, Mao H, Curran WJ, Liu T, 2016. 3D transrectal ultrasound (TRUS) prostate segmentation based on optimal feature learning framework, in: Medical Imaging 2016: Image Processing, International Society for Optics and Photonics. SPIE. pp. 654–660.

Yang X, Yu L, Wu L, Wang Y, Ni D, Qin J, Heng PA, 2017. Finegrained recurrent neural networks for automatic prostate segmentation in ultrasound images, in: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, AAAI Press. p. 1633–1639.

Yun S, Park J, Lee K, Shin J, 2020. Regularizing class-wise predictions via self-knowledge distillation, in: The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Zeng Q, Samei G, Karimi D, Kesch C, Mahdavi SS, Abolmaesumi P, Salcudean SE, 2018. Prostate segmentation in transrectal ultrasound using magnetic resonance imaging priors. International Journal of Computer Assisted Radiology and Surgery 13, 749–757. [PubMed: 29589259]

Zhan Y, Shen D, 2006. Deformable segmentation of 3-d ultrasound prostate images using statistical texture matching method. IEEE Transactions on Medical Imaging 25, 256–272. [PubMed: 16524083]

Zhou K, Yang Y, Hospedales T, Xiang T, 2020. Learning to generate novel domains for domain generalization, in: Vedaldi A, Bischof H, Brox T, Frahm JM (Eds.), Computer Vision - ECCV 2020, pp. 561–578.

Zhou Z, Rahman Siddiquee MM, Tajbakhsh N, Liang J, 2018. Unet++: A nested u-net architecture for medical image segmentation, in: Deep Learning in Medical Image Analysis and Multimodal Learning for clinical Decision Support, pp. 3–11. [PubMed: 32613207]
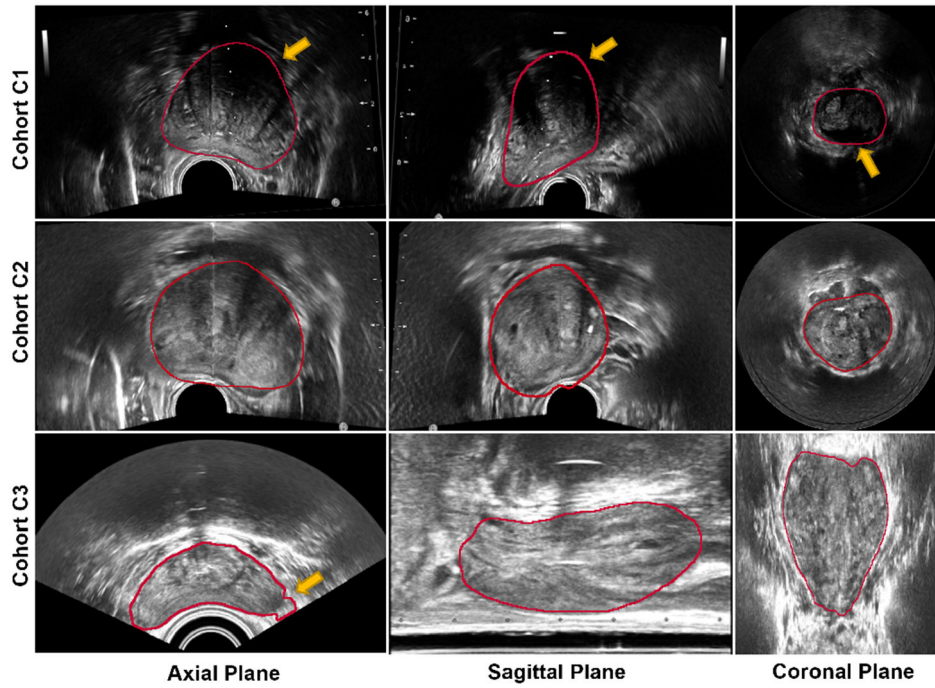
Fig. 1:
Example of prostate outlines (red) on ultrasound images acquired at three different institutions. Note the large variations in prostate shape, contrast, field of view, and the presence of artifacts (e.g., inhomogeneous intensity distributions). Moreover, the prostate boundary is not always clearly visible or easily distinguishable from the neighboring structures (orange arrows).
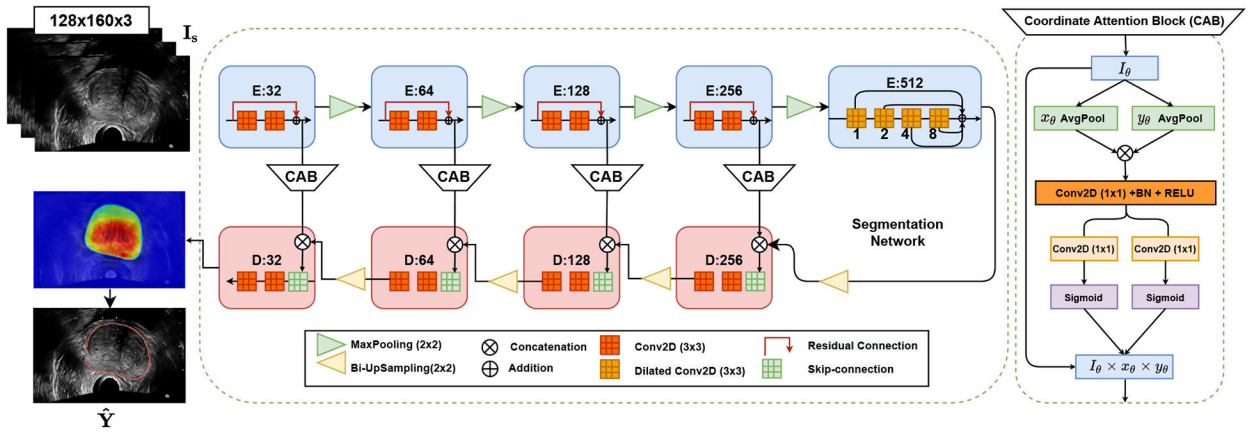
Fig. 2:
Flowchart of the dilated Residual UNet with coordination attention block and 2.5D input (128×160 of three neighboring slices used to provide spatial context). The output of the model is the probability map of the prostate segmented in the central slice. The attention blocks assist the model in reducing uncertainty in prostate borders, which are well known to be challenging to segment due to a lack of distinct borders. $\mathbf{I}_s$ is the input data. E and D refers to encoder and decoder blocks. $\widehat{\mathbf{Y}}$ is the predicted segmentation output.
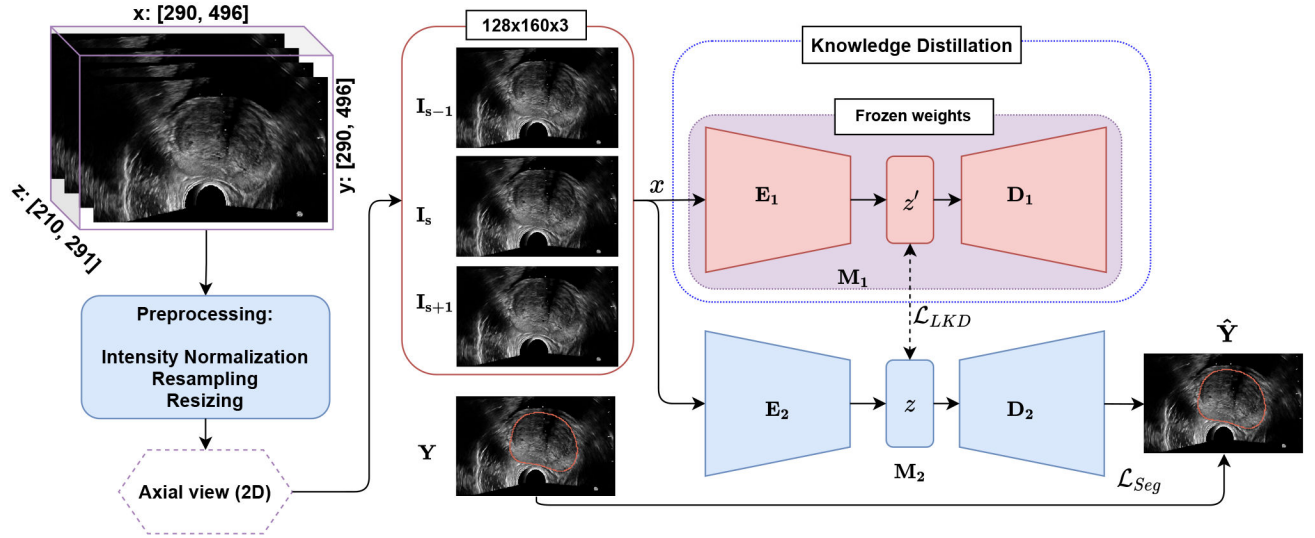
Fig. 3:
The overall pipeline for continual prostate gland segmentation. The segmentation model $\mathbf{M}_s$ is trained on cohort C1 in the first step, then the model is finetuned on cohorts C2 and C3. During optimization, the consistency of the features latent space of $\mathbf{M}_s$ and $\mathbf{M}_{t_1}$ is maintained by including the L2 norm loss (eq. 2). The same steps are repeated for finetuning the model trained on $\mathbf{M}_s$ and $\mathbf{M}_{t_1}$ for $\mathbf{M}_{t_2}$ model. $\mathbf{E}_1$, $\mathbf{E}_2$, $\mathbf{D}_1$, $\mathbf{D}_2$ are the encoders and decoders and z and z' are the latent space features for the source and target models. $\mathbf{Y}$ is the ground-truth and $\widehat{\mathbf{Y}}$ is the predicted segmentation output.
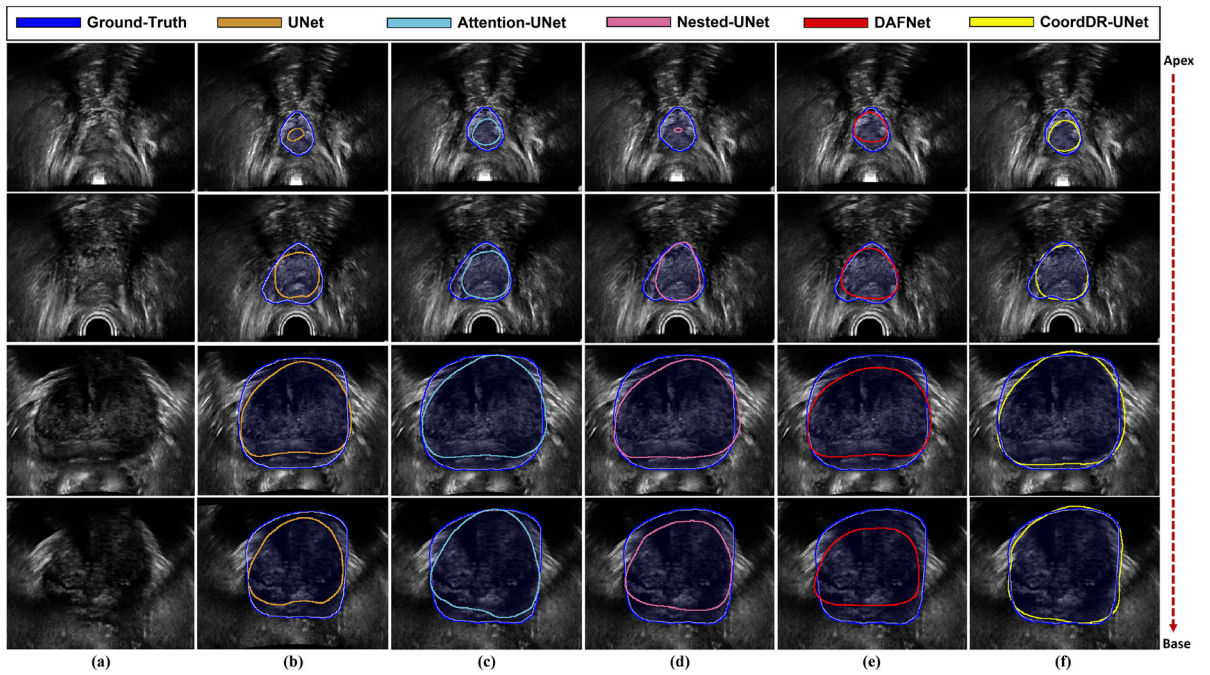
Fig. 4:

Visual comparison of prostate segmentation results produced by different methods for a sample patient in cohort C1-test. From left to right are the input TRUS slices (column (a)), the UNet 2.5D predictions (column (b)), Attention-UNet 2.5D (column c), Nested-UNet 2.5D (column (d)), DAFNet (Wang et al., 2019) (column (e)) and our proposed CoordDR-UNet 2.5D (column (f) for different slices from the apex (top row) to base (bottom row). The blue contours show the ground-truth segmentation outlined by an expert urologist.
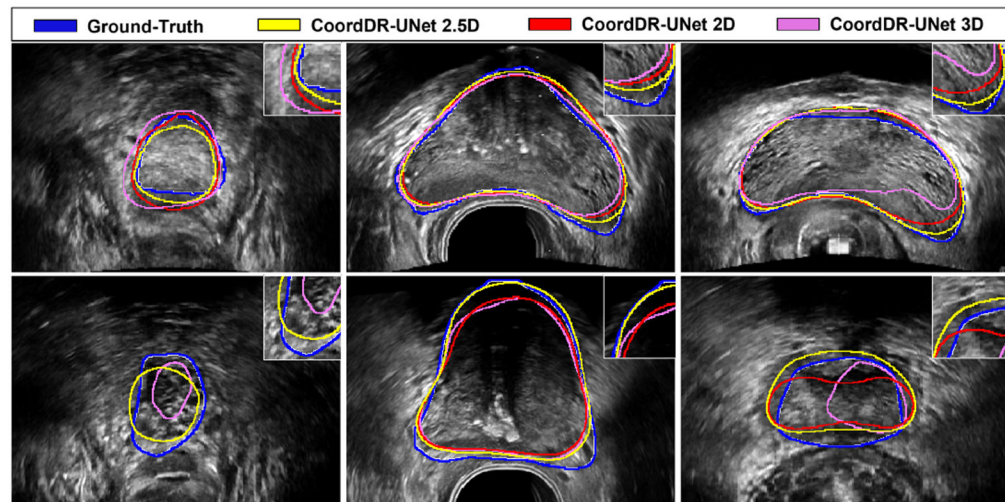
Fig. 5:
Visual comparison of CoordDR-UNet segmentation output with 2D, 3D, and 2.5D as the input. Each row shows apex, mid-gland, and base slices from different subjects from the cohort C1-test dataset.
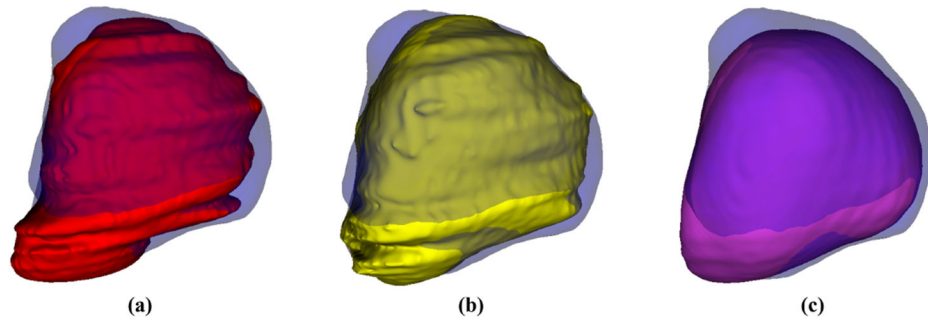
Fig. 6:

3D visualization of the segmentation results in a TRUS volume. CoordDR-UNet output with (a) 2D, (b) 2.5D and (c) 3D representations. The ground-truth mask is shown on a blue surface.
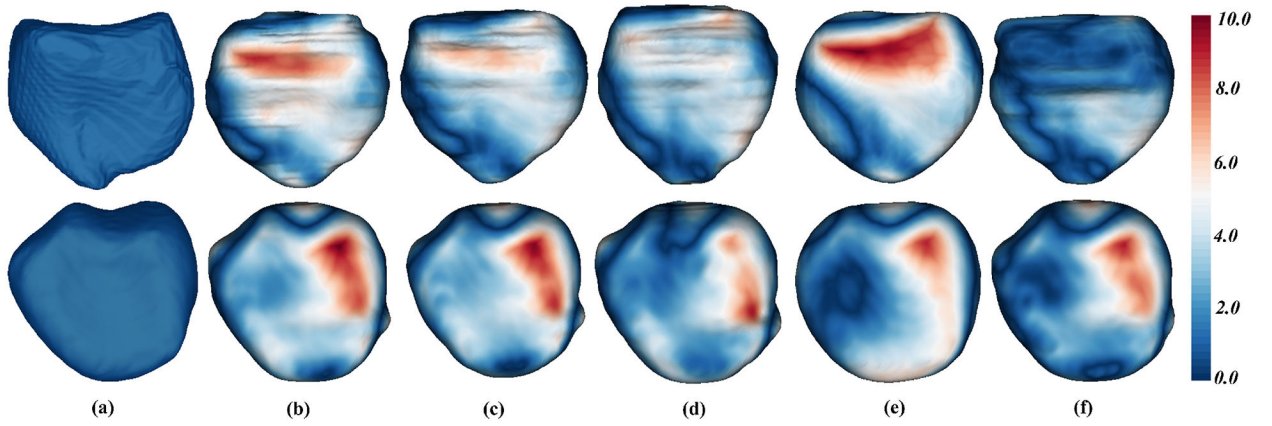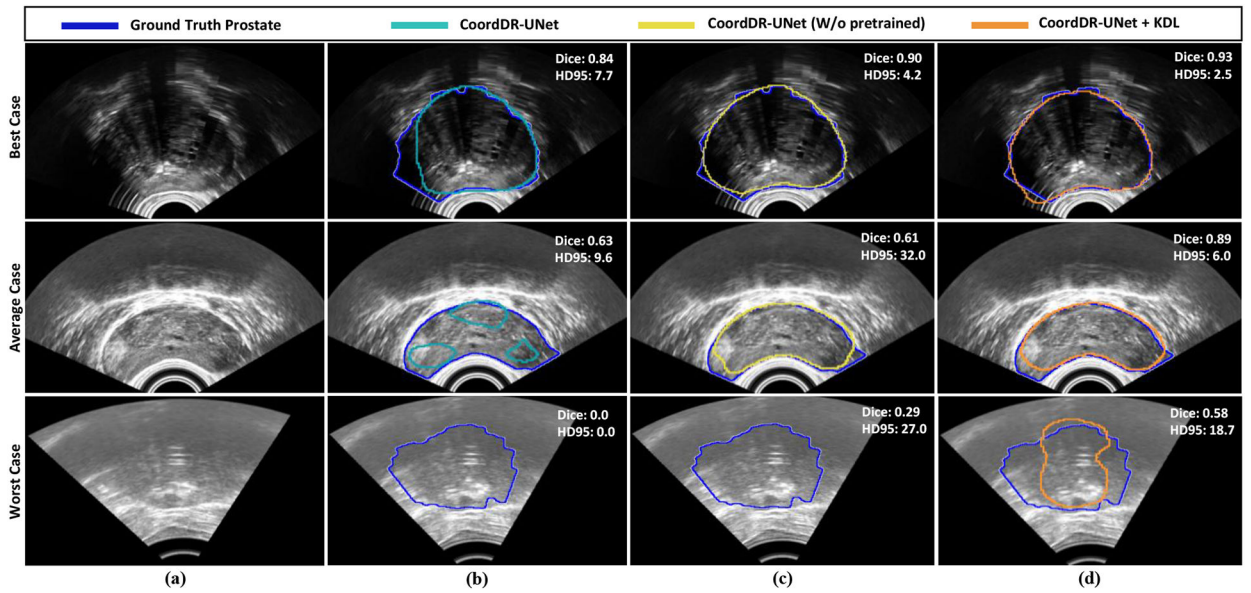
Fig. 7:
3D visualization of the surface distance (in voxel) between segmented surface and ground truth. Each row shows one subject. Different colors represent different surface distances. From left to right are (a) ground-truth, (b) UNet 2.5D, (c) Attention-UNet 2.5D, (d) Nested-UNet 2.5D, (e) DAFNet3D (Wang et al., 2019) and (f) our proposed CoordDR-UNet 2.5D. Our method consistently performs well on the whole prostate surface.

(a) Cohort C2-test
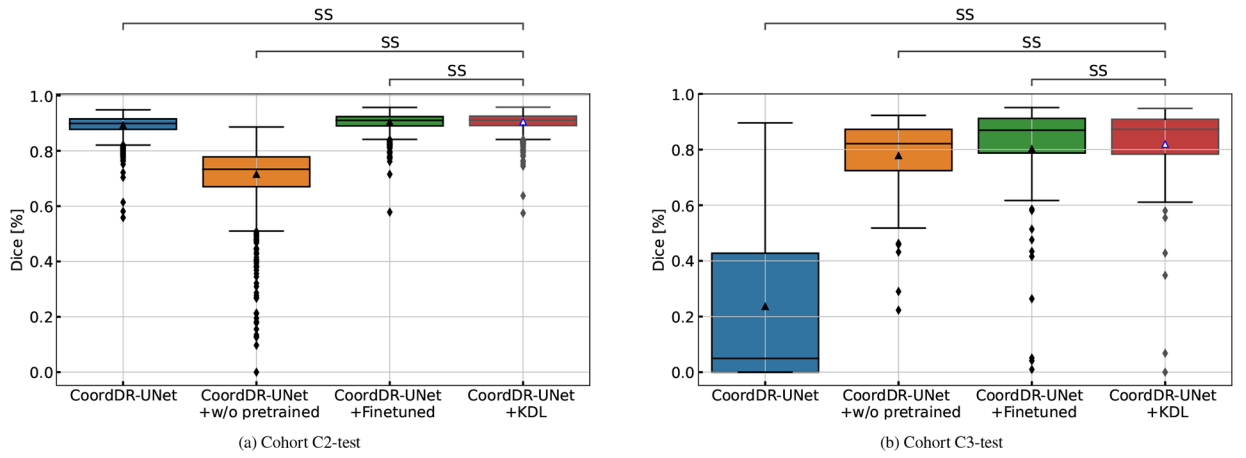
(b) Cohort C3-test

Fig. 8:

Box plots of Dice measure for the CoordDR-UNet, CoordDR-UNet + W/o pretrained, CoordDR-UNet with standard finetuning, and CoordDR-UNet + KDL segmentation approaches of cohort C2-test and C3-test. SS: statistically significant (P ≤ 0.05), NS: not significant (P > 0.05).
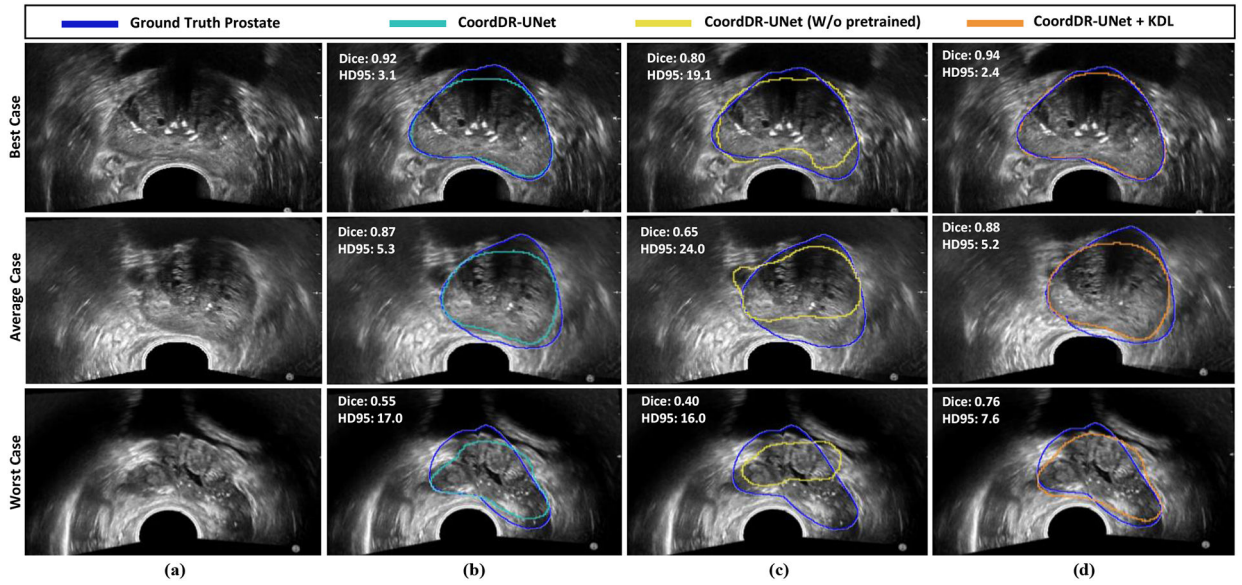
Fig. 9:
Visual comparison of segmentation results produced by different methods for three patients (best, average, worst) in cohort C2-test. From left to right are the 2D TRUS slices (column (a)), CoordDR-UNet 2.5D prediction without finetuning (column (b)), CoordDR-UNet 2.5D trained on ten TRUS images without any pretraining (column (c)), and our proposed CoordDR-UNet 2.5D finetuned on ten TRUS images with knowledge distillation loss (column (d)). Dice and HD95 scores are shown for each patient.

Fig. 10:

Visual comparison of segmentation results produced by different methods for three patients (best, average, worst) in cohort C3-test. From left to right are the 2D TRUS slices (column (a)), CoordDR-UNet 2.5D direct prediction without finetuning (column (b)), CoordDR-UNet 2.5D trained on ten TRUS images without any pretraining (column (c)), and our proposed CoordDR-UNet 2.5D finetuned on ten TRUS images with knowledge distillation loss (column (d)). Dice and HD95 scores are also shown for each patient.
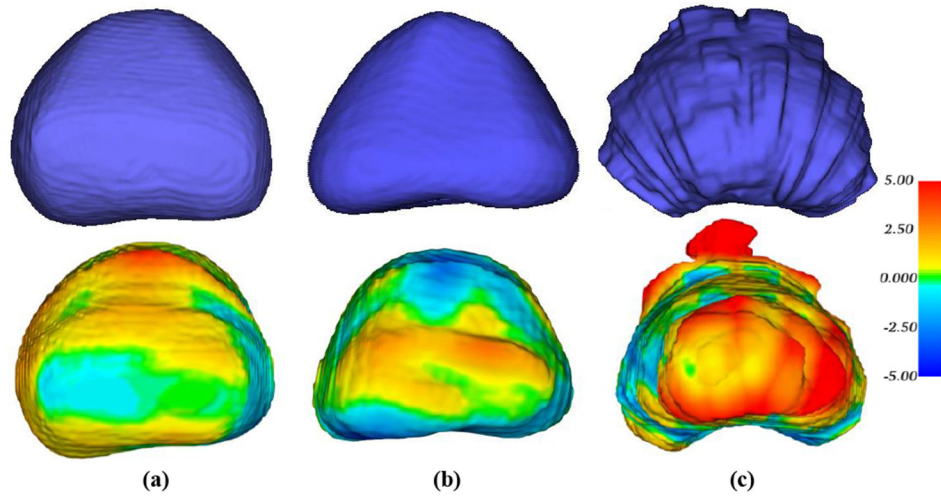
Fig. 11:

3D visualization of the surface distance (in voxel) between segmented prostate and ground truth. Different colors represent different surface distances. The top row shows the ground-truth surface masks for three different patients from cohort C1-test (a), C2-test (c), and C3-Test (b). The bottom row shows the computed HD surface distance between CoordDR-UNet 2.5D predictions and ground-truth.
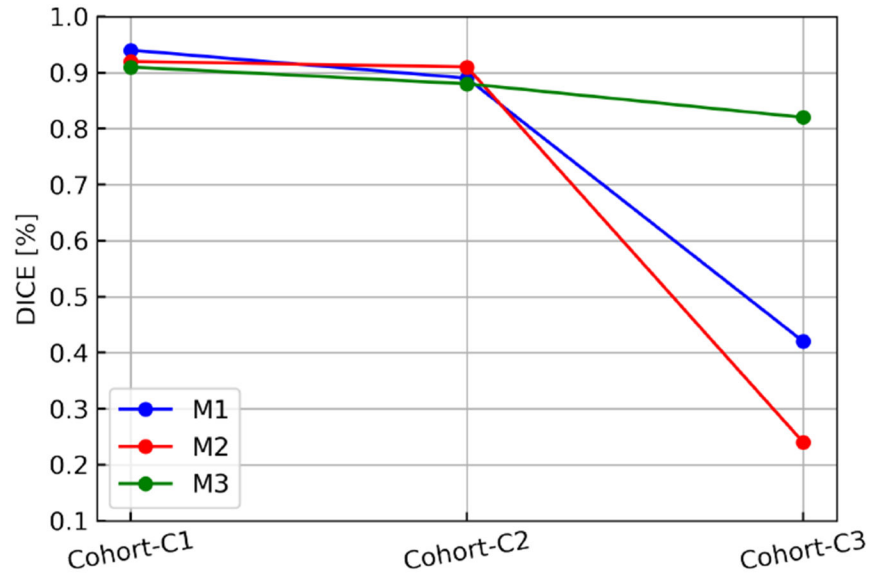
Fig. 12:
Continual segmentation results for domain generalization. $\mathbf{M}_s$: model trained using the train data from cohort C1, $\mathbf{M}_{t_1}$: model $\mathbf{M}_s$, finetuned using 10 TRUS images from cohort C2, and $\mathbf{M}_{t_2}$: model $\mathbf{M}_{t_1}$, finetuned using 10 TRUS images from cohort C3.
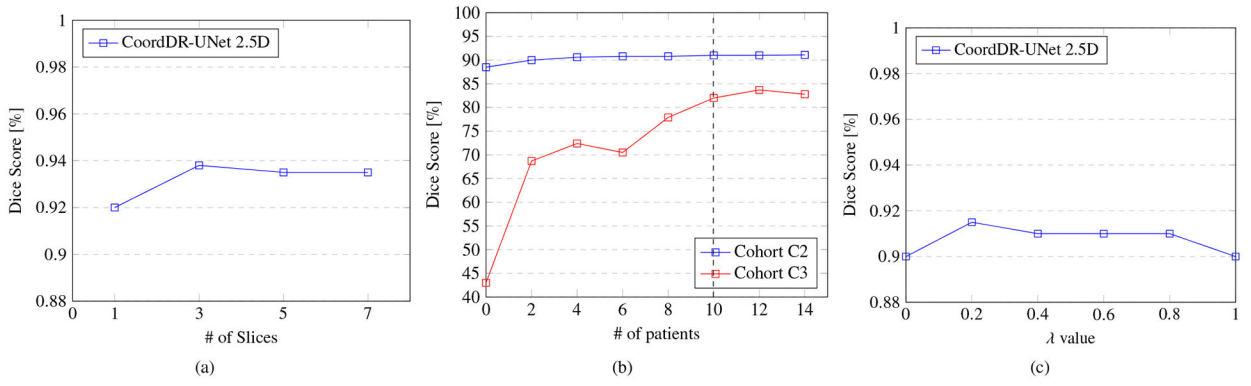
Fig. 13:
(a) CoordDR-UNet 2.5 performance with different number of slices on cohort C1-test set.
(b) CoordDR-UNet + LKD finetuning performance with different number of cases on cohort
C2-test and C3-test data. (c) CoordDR-UNet + LKD performance with different λ values on
cohort C2-test data.

Scanner specifications and demographic breakdown of the data included in our study.

Cohort	C1	C2	C3
Source	In house	Public	External
Number of Patients	764 (training), 190 (testing)	10 (training), 1,151 (testing)	10 (training), 96 (testing)
Number of Images	802 (training), 220 (testing)	10 (training), 1,751 (testing)	10 (training), 96 (testing)
Manufacturer	Eigen Inc.	Eigen Inc.	Hitachi
Scanner	Artemis	Artemis	HI VISION Preirus
Probe	Hitachi/Noblus/Hitachi C41V	Hitachi/Noblus/Hitachi C41V	Hitachi C41L47RP
Probe Type	End-fire	End-fire	Side-fire
Transmit Frequency (MHz)	7.5	7.5, 10	3.5
Frame Rate (Hz)	9	9	9
Average Pixel Spacing (mm)	[0.11, 0.52]	[0.21, 0.55]	[0.25]
Slice Thickness (mm)	[0.11, 0.52]	[0.21, 0.55]	[0.25]
Matrix Size (range)	290×290 - 496×496	342×342 - 452×452	303× 495
Number of Frames (range)	210 - 291	226 - 290	50 - 120

Table 2:

Quantitative comparison results (*mean*(±*std*)) between the proposed method and other segmentation methods for cohort C1-test dataset. The best results are highlighted in bold. Paired Student's t-tests showed statistical significance (P 0.05) when comparing the methods with the baseline method 2D UNet. DAFNet: Deep Attentive Features Network, DR-UNet: Dilated-Residual UNet, CoordDR-UNet: Coordination Dilated-Residual UNet, HD95: 95th percentile of the Hausdorff distances.

		Evaluation on cohort C1-test (n=220)		
Methods	**Input**	**Dice**	**HD95 [*mm*]**	**Sensitivity**
UNet (Ronneberger et al., 2015)	2D	0.90 (±0.03)	4.03 (±3.52)	0.94 (±0.04)
CoordDR-UNet (proposed)		0.92 (±0.03)	3.29 (±1.45)	0.92 (±0.05)
UNet (Ronneberger et al., 2015)	3D	0.90 (±0.03)	5.21 (±3.52)	0.92 (±0.01)
DAFNet (Wang et al., 2019)		0.92 (±0.03)	2.87 (±1.27)	0.90 (±0.04)
CoordDR-UNet (proposed)		0.92 (±0.04)	3.00 (±1.46)	0.92 (±0.07)
UNet (Ronneberger et al., 2015)	2.5D	0.91 (±0.07)	3.00 (±1.30)	0.91 (±0.08)
Att-UNet (Schlemper et al., 2019)		0.92 (±0.02)	2.75 (±1.16)	0.93 (±0.04)
Nested-Unet (Zhou et al., 2018)		0.91 (±0.04)	3.44 (±1.55)	0.93 (±0.06)
DR-UNet (Vesal et al., 2021)		0.93 (±0.02)	2.34 (±0.92)	0.94 (±0.03)
CoordDR-UNet (proposed)		**0.94 (±0.03)**	**2.29 (±1.45)**	**0.94 (±0.05)**

Table 3:

Quantitative comparison results (*mean*(±*std*)) for different finetuning strategies shown for cohort C2-test and C3-test. The best results are highlighted in bold. Paired Student's t-tests showed statistical significance (P 0.05) when comparing the CoordDR-UNet + KDL with no pretraining and direct prediction models. Bolded entries represent the best metric in each test set.

Methods		Dice	HD95 [*mm*]	Sensitivity
Evaluation on cohort C2-test (n= 1,751)				
CoordDR-UNet	exp1	0.89 (±0.03)	4.03 (±1.62)	0.84 (±0.06)
CoordDR-UNet + w/o pretrained	exp2	0.71 (±0.09)	14.8 (±4.60)	0.77 (±0.12)
CoordDR-UNet + Finetuning	exp3	0.90 (±0.03)	3.80 (±1.60)	0.87 (±0.05)
CoordDR-UNet + KDL	exp4	**0.91 (±0.03)**	**3.69 (±1.49)**	**0.88 (±0.05)**
Evaluation on cohort C3-test (n=96)				
CoordDR-UNet	exp1	0.24 (±0.29)	14.14 (±11.83)	0.18 (±0.25)
CoordDR-UNet + w/o pretrained	exp2	0.78 (±0.14)	10.23 (±6.59)	0.76 (±0.18)
CoordDR-UNet + Finetuning	exp3	0.80 (±0.19)	7.30 (±6.21)	0.73 (±0.10)
CoordDR-UNet + KDL	exp4	**0.82 (±0.16)**	**7.13 (±6.25)**	**0.76 (±0.19)**