Published in final edited form as:
Proc SPIE Int Soc Opt Eng. 2021 February ; 11598: . doi:10.1117/12.2581963.

Three-dimensional prostate CT segmentation through fine-tuning of a pre-trained neural network using no reference labeling

Kayla Caughlin[a],
Maysam Shahedi[a],
Jonathan E. Shoag[b,c],
Christopher Barbieri[b],
Daniel Margolis[d],
Baowei Fei[a,e,f,*]

[a]Department of Bioengineering, The University of Texas at Dallas, Richardson, TX

[b]Department of Urology, Weill Cornell Medicine, New York, NY

[c]Department of Urology, University Hospitals Medical Center, Case Western Reserve University, Cleveland, Ohio

[d]Department of Radiology, Weill Cornell Medicine, New York, NY

[e]Advanced Imaging Research Center, University of Texas Southwestern Medical Center, Dallas, TX

[f]Department of Radiology, University of Texas Southwestern Medical Center, Dallas, TX

Abstract

Accurate segmentation of the prostate on computed tomography (CT) has many diagnostic and therapeutic applications. However, manual segmentation is time-consuming and suffers from high inter- and intra-observer variability. Computer-assisted approaches are useful to speed up the process and increase the reproducibility of the segmentation. Deep learning-based segmentation methods have shown potential for quick and accurate segmentation of the prostate on CT images. However, difficulties in obtaining manual, expert segmentations on a large quantity of images limit further progress. Thus, we proposed an approach to train a base model on a small, manually-labeled dataset and fine-tuned the model using unannotated images from a large dataset without any manual segmentation. The datasets used for pre-training and fine-tuning the base model have been acquired in different centers with different CT scanners and imaging parameters. Our fine-tuning method increased the validation and testing Dice scores. A paired, two-tailed t-test shows a significant change in test score ($p = 0.017$) demonstrating that unannotated images can be used to increase the performance of automated segmentation models.

[*] bfei@utdallas.edu, Website: https://fei-lab.org.

DISCLOSURE
The authors have no relevant financial interests in this article and no potential conflicts of interest to disclose.

Keywords

Image segmentation; deep learning; pre-trained neural networks; fine tuning; prostate; computed tomography (CT)

1. INTRODUCTION

Recently, a variety of methods have been used with the goal of automated or semi-automated prostate segmentation[1–5]. Some of the previously presented prostate CT image segmentation methods that produced promising results used intra-patient overlap in the training and test sets[3]. More current methods do not rely on images from each patient in both the training and test sets, but instead, guide the model using multiple references or user annotations. For example, Shahedi, et al. requires the training images to have two expert segmentations for best results[2]. Another method required user input of bounding box information and user selection of twelve points on the prostate border[1]. Shahedi, et al. experiments with a range of user input points, with results peaking with 15–20 points[5]. While the number of prostate CT images currently available is large, datasets with expert manual segmentation labels are small. Thus, new research is trending towards improving the results on small labeled datasets by including data augmentation and/or some level of user input. However, for clinical applications of deep learning-based segmentation with a high quantity of images (such as automated volume tracking over time), fully automated methods that also generalize to varied clinical imaging specifications and locations are needed. Fully automated methods are also advantageous in settings such as volume tracking where relative differences in volume over time need to be quantified. In this setting, inter- and intra- observer variability in manual approaches could obscure small trends over time.

In this work, we focused on increasing the performance of automated segmentation methods by using a fine-tuning approach with weakly-labeled images. We first trained a base model using a small dataset with expert segmentation labels. We then used an unsupervised, iterative, fine-tuning method to gradually introduce new images from an independent dataset with varied image specifications and no expert segmentation. While previous work has used weakly-labeled images in prostate magnetic resonance imaging (MRI) segmentation, weak labels were generated using a sampling of points from a manual segmentation[4]. A concept similar to ours was used in Bai, et al., with a base network predicting a segmentation for unlabeled images[6]. However, Bai, et al. applied the concept to cardiac MR segmentation in a 2D network and automatic segmentations are improved using a conditional random field[6]. In addition to using unlabeled images, we used large input volumes to allow the use of images with different voxel sizes and prevent the method from relying on precise localization of prostate landmarks by a trained user.

2. METHODS

2.1 Data

Our data consists of two independent datasets each containing abdominal CT scans, including images with artifacts caused by brachytherapy seeds and metallic implants. The first dataset (referred to as the base dataset) contains 92 CT scans of 92 different prostate cancer patients with manual, expert segmentation for each image, and a voxel size of 0.977×0.977×4.25 mm^3. The second dataset (referred to as the fine-tuning dataset) contains over 300 abdominal CT scans but lacks any manual segmentation data, includes multiple scans of each patient over several years, has pixel spacing ranging from 0.660 mm to 0.977 mm, and slice thickness ranging from 2.50 mm to 3.75 mm. We used 60 annotated images for training the base model and up to 80 unannotated images for fine-tuning. We used 10 annotated images for validation during training and fine-tuning. Twenty two annotated images were left reserved for final tests.

2.2 Preprocessing

We used a generously large bounding box size of 128×128×17 voxels (increasing the cropped image volume by approximately 100% from previous work in[2]) to ensure that the prostate was fully contained within the cropped volume for both datasets. The fine-tuning dataset images tended to have a smaller voxel size and thus required a larger bounding box than the base dataset. By using a large bounding box, we eliminated the need to precisely locate the base and apex of the prostate and enabled the model to adapt to differences in datasets without additional pre-processing such as image resizing. In addition to requiring a larger bounding box, the Hounsfield units (HU) in the fine-tuning dataset were corrupted and image intensities required modification. Based on differences between the values in the fine-tuning dataset images and the standard reference values for air, water, and urine, we approximated standard HU by subtracting 1022 from each voxel intensity. After modifying the HU on the fine-tuning dataset only, to increase background consistency across the data, HU values of both data sets were truncated to the −69 to +165 range (the observed HU range for prostate tissue[2]).

2.3 Network Architecture

As shown in Figure 1, we used a base training stage and a fine-tuning stage. We used a standard four-level 3D U-Net architecture[7] to train the base model. Each of the four levels was composed of multiple convolutional layers. The network model was similar to what we used in our previous study[2]. We automatically generated weak labels for the fine-tuning dataset using the base model and its subsequent updates. The last 20 layers in the model were frozen and the learning rate was set to 0.7 during fine-tuning to prevent imperfections in the weakly labeled images from causing large alterations in the model. In addition, we gradually incorporated the weakly labeled images to the base training set and frequently refreshed the weak labels to iteratively improve the quality of the weak labels during fine-tuning.

2.4 Implementation details

We used TensorFlow framework[8] and Keras libraries to implement both the base model and the fine-tuning process. The 92 base images were randomly divided into training (60), validation (10), and testing (22) groups. We used horizontal reflections to augment the base training data, resulting in a base training set of 120 images. The validation set was not augmented during training. The validation and testing sets were not altered between the base and fine-tuning stages. During fine-tuning, 80 images from the fine-tuning dataset were gradually incorporated into the training set for a total of 200 training images at the conclusion of fine-tuning. In both stages, the batch size was one and the Adadelta optimizer[9] was used with a loss function based on Dice similarity coefficient[10] introduced in[11] (hereafter called "soft Dice") to optimize the training. The initial learning rate was set to the default during base training and then set to 0.7 in the fine-tuning process.

2.5 Evaluation

We evaluated the fine-tuning method using Dice coefficient as the segmentation error metric. We compared the network output for the validation and testing images against their corresponding manual segmentations. We calculated the final probability maps by pixel-wise averaging the probability map of each test image and its reflected version. The average probability map was converted to a binary map using a 50% threshold level for measuring the Dice coefficient.

3. RESULTS

2.1 Training

We trained the base model for a total of 300 epochs and selected the best model based on the validation loss. Figure 2A shows the minimum validation loss occurred at epoch 217 with a value of 24.1%. The corresponding training loss was 16.4%. The growing gap between the training and validation losses after epoch 217 suggests that the model is overfitting to the training set and is unlikely to learn a better representation with more base training time.

Figure 2B shows the optimal validation score from the 10 epochs of fine-tuning for each additional image added to the base training set. The validation score for the base model (no additional images) was 75.9% and increased by 2.1% to a maximum of 78.0% at additional image 53.

2.2 Testing

The Dice score on the testing set for the base model (selected at base training epoch 217 based on validation score) was 77.37%. The Dice score on the testing set for the fine-tuned model (selected after the addition of 53 images based on validation score) was 78.77%, showing an overall improvement of 1.40% over the base model. Using a paired, two-tailed t-test, the change in test score is statistically significant ($p = 0.017$). Individually, 15/22 images showed an improvement in Dice score, while 7/22 images showed a decrease in Dice score. The greatest decrease in Dice score was 1.64%, while the greatest increase in Dice score was 10.29%. A 3D visualization to compare the results is shown in Figure 3. Each row in Figure 3 corresponds to a single testing image, with a total of four images presented. The predicted base segmentations are shown in yellow, while the fine-tuned predictions are shown in purple. The manual segmentations are shown in blue. The image that showed the greatest increase was also the image with the lowest base Dice score of 56.0%, while the image with the greatest decrease had a base Dice score over 80.0% (see Figure 3, rows 1 and 4). Figure 4 shows a qualitative comparison of model performance at the base, mid-gland, and apex slices for the same four images shown in Figure 3.

4. DISCUSSION AND CONCLUSIONS

We developed a fine-tuning method for a pre-trained deep learning segmentation model using a dataset with no manual segmentation labels. We automatically generated segmentation labels for the fine-tuning data using the model under fine-tuning and its weakly-labeled images from a dissimilar dataset to improve the performance of a deep learning-based 3D prostate CT segmentation on a base dataset. We used the pre-trained base model to automatically generate weak labels for fine-tuning with no user interaction such as selection of points on the prostate boundary and no sampling of points from a manual segmentation label as used in some of the previous studies[1,4]. The fine-tuning procedure significantly improved the Dice score on the testing set by 1.4% in average. We achieved an improvement in Dice score for 15/22 test subjects, with the maximum improvement on a single image over 10%. The overall Dice score for the test set using the fine-tuned network was about 79%. Other results for automated prostate segmentation in CT images have presented a range of higher testing performance[2,12,13]. Our base model may benefit from further tuning of hyperparameters, or by implementing additional techniques, such as the cross-validation strategy and the weighted loss function successfully used for prostate segmentation in[12]. In addition, we have not post-processed the segmentation labels. We believe applying a post-processing step to our algorithm could improve the overall results. For example, in Figure 3, the small miss-classified regions at the prostate base side could be easily removed during post-processing. However, we consider this a preliminary study for assessing the potential of unsupervised methods for fine-tuning automated prostate segmentation models. Thus, we focus our evaluation on the increase in Dice score between our base model and our fine-tuned model and leave improvements in base model performance for future work.

4.1 Limitations

Our method achieved a maximum 2.1% increase in validation Dice score after adding 53 fine-tuning images. The decrease in validation score following image 53 (Figure 3) may indicate that improvement from weakly labeled images is capped when the additional image count approaches the manually labeled image count. However, Figure 3 shows that the score trends up again after image 68 and reaches within 0.13% of the best validation score by image 77. Thus, the fine-tuning method may be adjusting to the introduction of a more challenging weakly labeled image (e.g. an image with considerable distortion from an artifact) resulting in a lower-quality weak segmentation. The model may be able to adjust to the initial dip in performance and gradually recover by the addition of higher quality weakly-labeled images or through iteratively improving the weak label of the challenging image. However, the network may benefit from an additional constraint on the weakly labeled images, such as the size constraint used by a previous study[4].

4.2 Conclusions

We introduced a new training method for prostate segmentation using a pre-trained base model and introduction of unannotated images in the fine-tuning stage, resulting in an average improvement in test Dice coefficient. Additionally, we use a large bounding box, no user annotation of points on the prostate, and a diverse collection of training images collected with various imaging parameters. Our method is applicable to scenarios where a limited amount of manual segmentations exists, but in which the clinical setting requires a large quantity of images to be processed. While our method increased the average Dice score on the test set, additional improvements may be achieved by optimizing the number of frozen layers during fine-tuning and the number of additional images to include at each step of fine-tuning. We would also like to obtain manual segmentations for a subset of the fine-tuning dataset in order to measure if the method is able to improve the Dice score of both datasets simultaneously.

ACKNOWLEDGMENTS

This research was supported in part by the U.S. National Institutes of Health (NIH) grants (R01CA156775, R01CA204254, R01HL140325, and R21CA231911) and by the Cancer Prevention and Research Institute of Texas (CPRIT) grant RP190588.

REFERENCES

[1]. Shahedi M, Ma L, Guo R, Zhang G, Schuster DM, Nieh PT, Master V. v., Halicek M and Fei B, “A semiautomatic algorithm for three-dimensional segmentation of the prostate on CT images using shape and local texture characteristics,” Medical Imaging 2018: Image-Guided Procedures, Robotic Interventions, and Modeling, 1057616 (2018) 33.

[2]. Shahedi M, Halicek M, Dormer JD, Schuster DM and Fei B, “Deep learning-based three-dimensional segmentation of the prostate on computed tomography images,” Journal of Medical Imaging (2019).

[3]. Li W, Liao S, Feng Q, Chen W and Shen D, “Learning image context for segmentation of the prostate in CT-guided radiotherapy,” Physics in Medicine and Biology (2012).

[4]. Kervadec H, Dolz J, Tang M, Granger E, Boykov Y and ben Ayed I, “Constrained-CNN losses for weakly supervised segmentation,” Medical Image Analysis (2019).

[5]. Shahedi M, Halicek M, Dormer JD and Fei B, “Incorporating minimal user input into deep learning based image segmentation,” Medical Imaging 2020: Image Processing, 1131313 (2020).

[6]. Bai W, Oktay O, Sinclair M, Suzuki H, Rajchl M, Tarroni G, Glocker B, King A, Matthews PM and Rueckert D, “Semi-supervised learning for network-based cardiac MR image segmentation,” Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (2017).

[7]. Ronneberger O, Fischer P and Brox T, “U-Net: Convolutional Networks for Biomedical Image Segmentation [2015; First paper exploring U-Net architecture.],” International Conference on Medical image computing and computer-assisted intervention (2015).

[8]. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, Devin M, Ghemawat S, Irving G, Isard M, Kudlur M, Levenberg J, Monga R, Moore S, Murray DG, Steiner B, Tucker P, Vasudevan V, Warden P, et al. , “TensorFlow: A system for large-scale machine learning,” Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2016 (2016).

[9]. Zeiler MD, “ADADELTA: an adaptive learning rate method” (2012). Available: arxiv:1212.5701.

[10]. Dice LR, “Measures of the Amount of Ecologic Association Between Species,” Ecology (1945).

[11]. Milletari F, Navab N and Ahmadi SA, “V-Net: Fully convolutional neural networks for volumetric medical image segmentation,” Proceedings - 2016 4th International Conference on 3D Vision, 3DV 2016 (2016).

[12]. Balagopal A, Kazemifar S, Nguyen D, Lin MH, Hannan R, Owrangi A and Jiang S, “Fully automated organ segmentation in male pelvic CT images,” Physics in Medicine and Biology (2018).

[13]. Kazemifar S, Balagopal A, Nguyen D, McGuire S, Hannan R, Jiang S and Owrangi A, “Segmentation of the prostate and organs at risk in male pelvic CT images using deep learning,” Biomedical Physics and Engineering Express (2018).

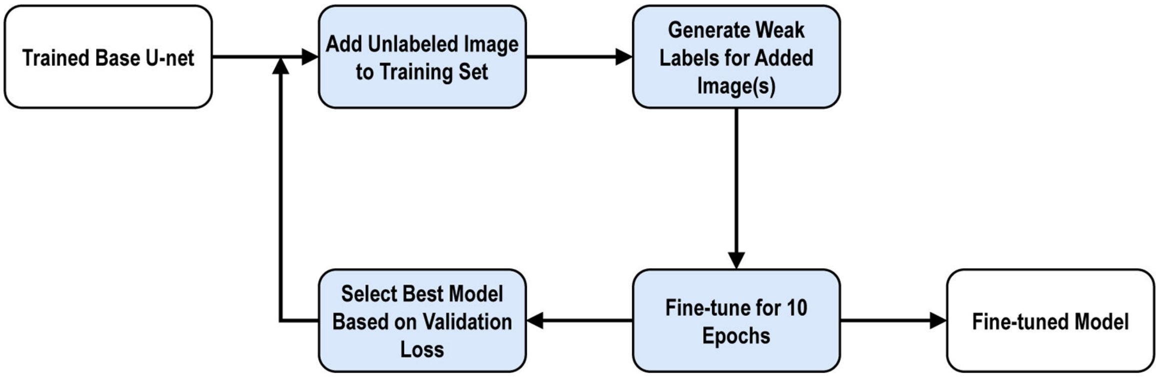
Trained Base U-net
Add Unlabeled Image to Training Set
Generate Weak Labels for Added Image(s)
Fine-tune for 10 Epochs
Select Best Model Based on Validation Loss
Fine-tuned Model

Figure 1:

Training block diagram for fine-tuning.

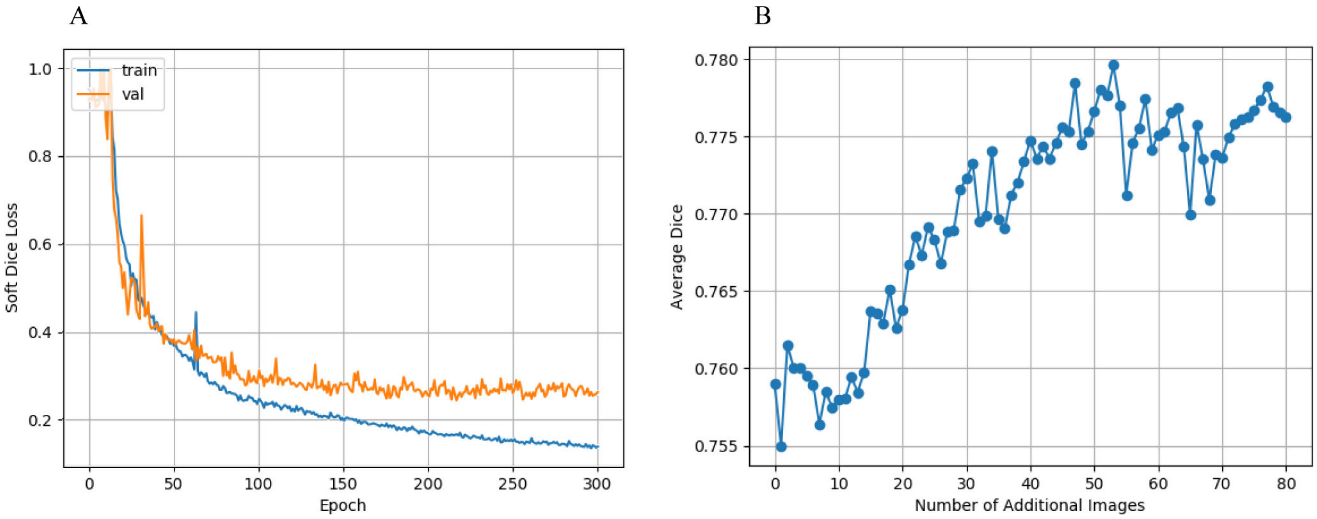

Figure 2:

Training curves of the neural networks. A: Base model training and validation loss curves. B: Fine-tuning validation Dice score per fine-tuning image addition.

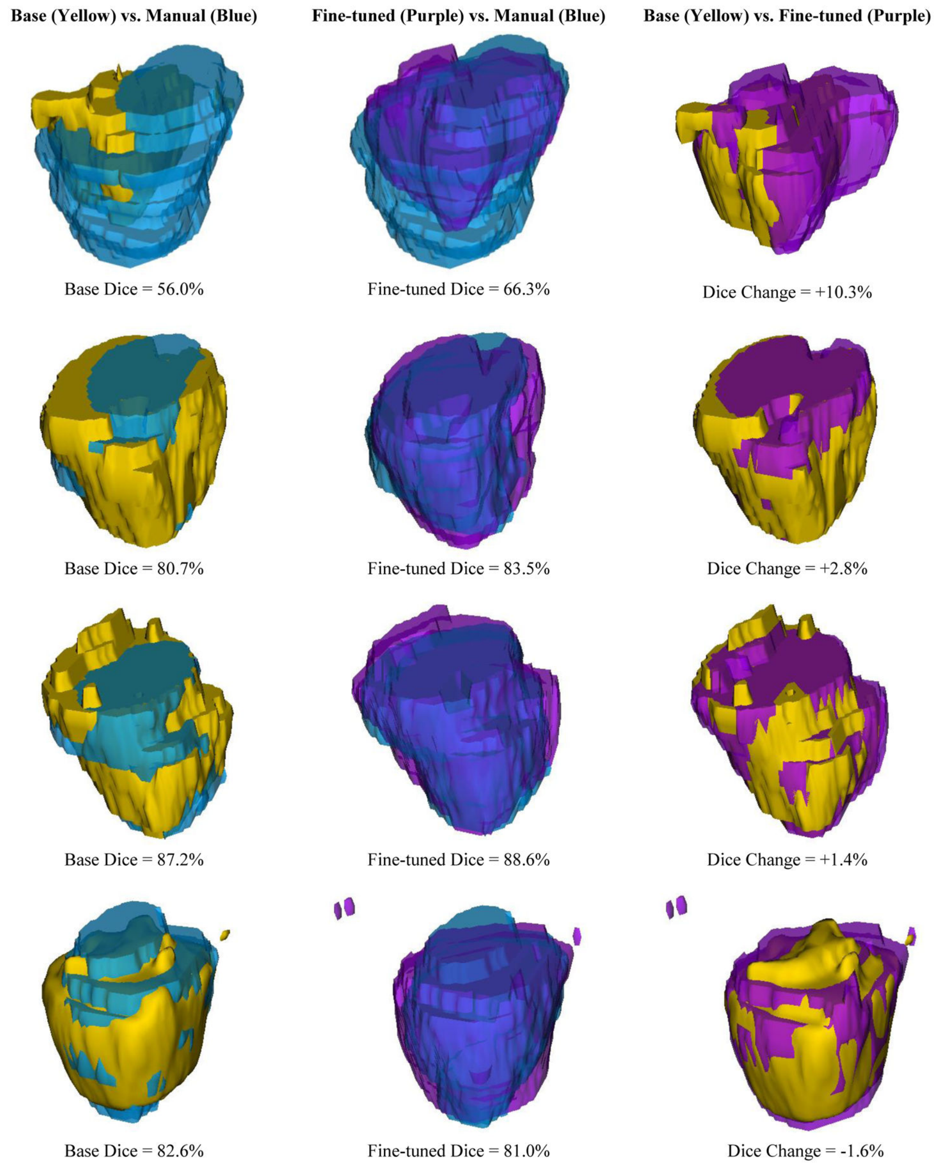

Figure 3:

3D visualization of four sample test images showing a qualitative comparison of base, fine-tuned, and manual segmentations.

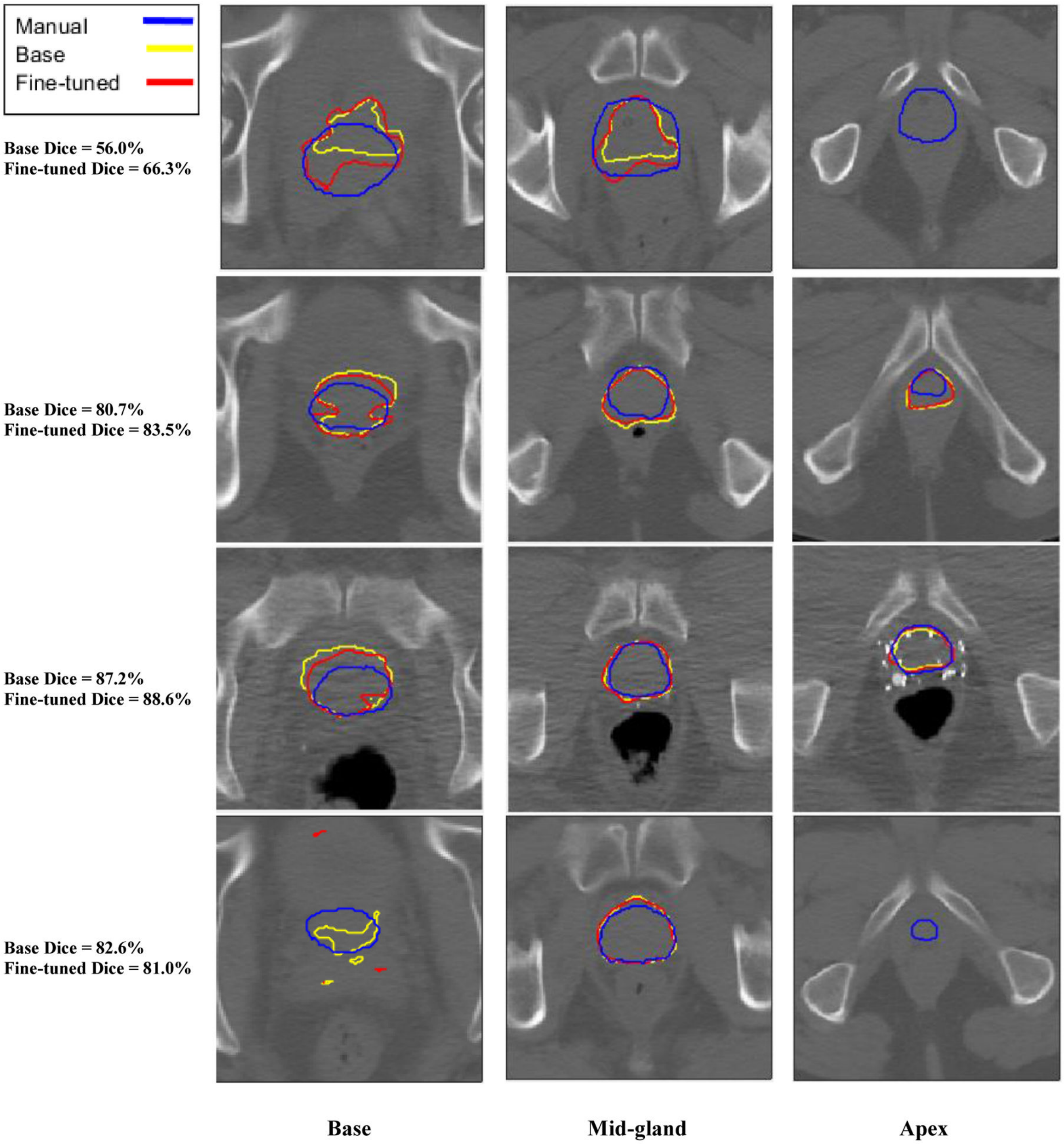

Figure 4:

Qualitative comparison of manual, base, and fine-tuned model segmentations at the base, mid-gland, and apex axial slices for four sample test images.