SUPPLEMENTARY MATERIALS

Dataset

The details of data processing are shown in Figure S1 and are as follows:

First, images are registered to align T2w and ADC images. Then, cropping based on CNN is applied to T2w images (crop a square region encompassing the entire prostate region for every T2w slice)[1,2]. Then, the identified square is propagated to the corresponding registered ADC slices. Finally, image augmentation based on moving Least Squares is applied to obtain a larger training dataset.

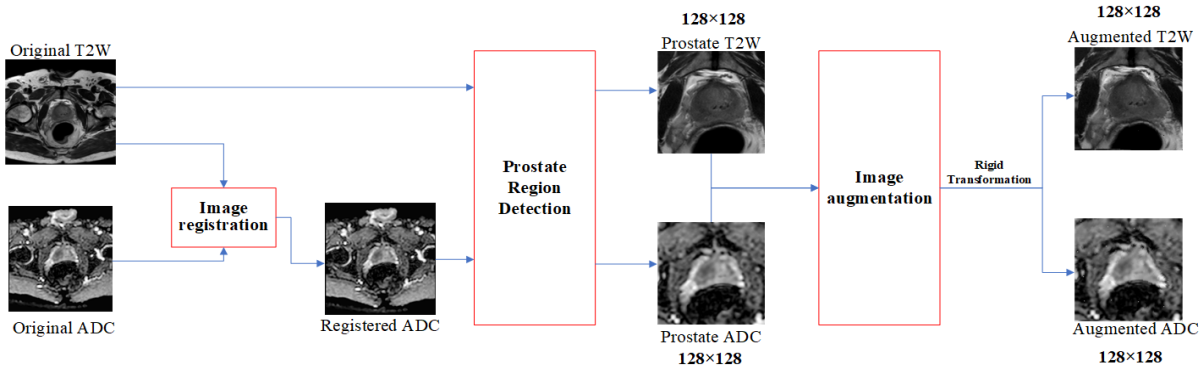

FIGURE S1. Details of data processing which consists of image registration, prostate region detection, and image augmentation.

Model training

Evaluation of model training

The specific training process is as follows:

First, we train the coarse segmentation component on the dataset to generate coarse prostate masks. Then, we concatenate the mp-MRI with the corresponding coarse prostate masks generated by the coarse segmentation component as the input to the classification component. We train the classification component to generate the PCa classification results and output the CRM. Finally, we fuse the cropped T2w images and the corresponding CRM refined by the classification component, and feed them to the fine segmentation component to train the fine segmentation component to deliver more accurate prostate masks.

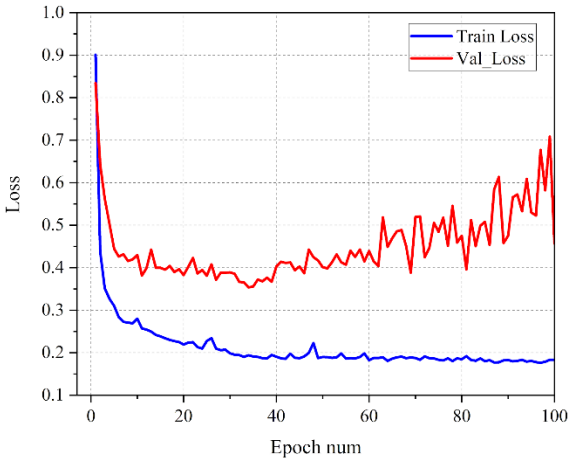

FIGURE S2. Learning curves of Coarse Segmentation component. The blue and red lines represent training and validation loss, respectively. The x axis represents the number of training epochs.

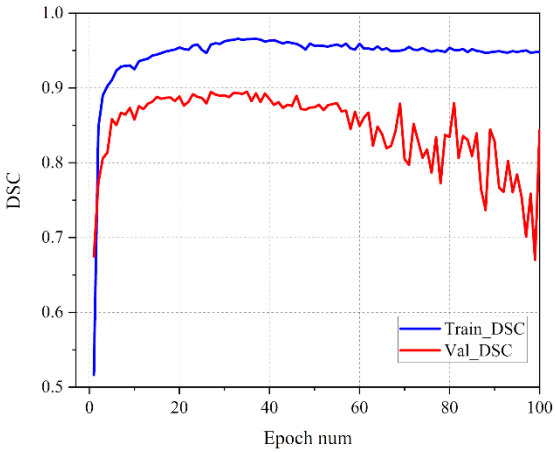

FIGURE S3. Learning curves of Coarse Segmentation component. The blue and red lines represent training and validation DSC, respectively. The x axis represents the number of training epochs.

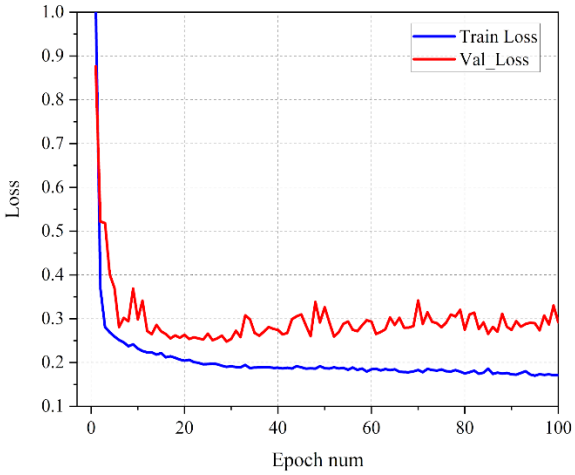

FIGURE S4. Learning curves of Fine Segmentation component. The blue and red lines represent training and validation loss, respectively. The x axis represents the number of training epochs.

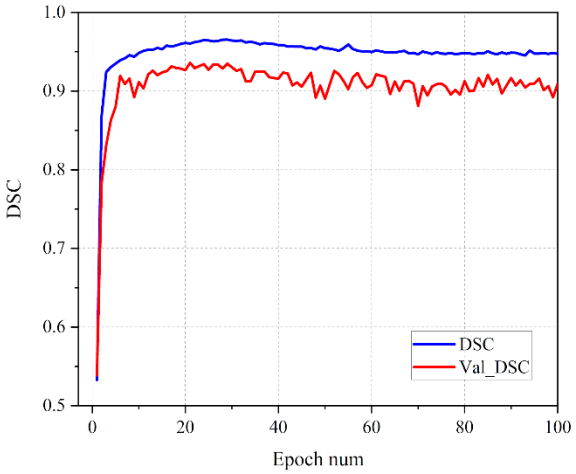

FIGURE S5. Learning curves of Fine Segmentation component. The blue and red lines represent training and validation DSC, respectively. The x axis represents the number of training epochs.

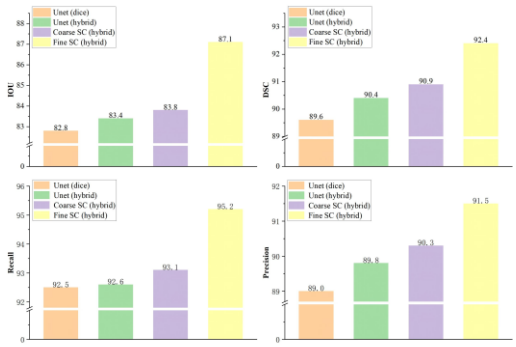

FIGURE S7. Comparison of quantitative metrics for different segmentation networks, including Unet with Dice and hybrid loss, coarse segmentation component (Coarse SC), and fine segmentation component (Fine SC) for center B. The loss is indicated in brackets, including Dice loss (Dice) and segmentation hybrid loss (hybrid). Both Coarse SC and Fine SC are based on a residual Unet with attention blocks.

Validation of the contribution of segmentation to classification

To further validate improvement of the performance of the network when coarse segmentation masks are included, we also compared heatmaps obtained by the classification component with or without using prostate masks, as shown in Figure S6. When using the coarse prostate masks, coincidence of the areas in heatmaps displaying the highest values with the areas where the prostate is located is improved.

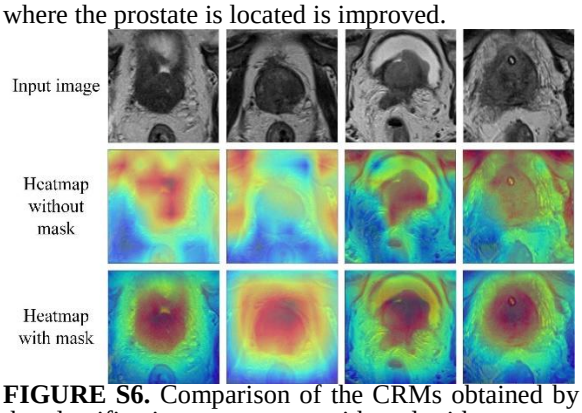

FIGURE S6. Comparison of the CRMs obtained by the classification component with and without usage of the coarse prostate masks. First row: input images. Second row: heatmaps obtained without using as an input the coarse prostate masks. Third row: heatmaps obtained using as an input the coarse prostate masks.

Segmentation results

Comparison of segmentation results of MC-DSCN and Unet

The statistical metrics corresponding to different models for center B are shown in Figure S8.

Comparison of segmentation results of MC-DSCN and Unet++

To demonstrate the effectiveness of the segmentation component in MC-DSCN, we also compared our segmentation network with Unet++. Comparison of quantitative metrics for different segmentation networks are shown in Figure S9 (for center A) and Figure S10 (for center B).

For the coarse segmentation, the quantitative metrics of IOU, DSC, and Recall were better when using Unet++. For the fine segmentation, the evaluation indicators of IOU, DSC, and Precision were better when using Unet++. Compared with Unet, Unet++ combines the simple features of the shallow layer with the abstract features of the deep layer, and integrates them through feature superposition. It has receptive fields of different sizes, so its performance is better than Unet. Moreover, we can notice that for the segmentation component, whether Unet or Unet++ is used, the performance of Fine SC is always better than that of Coarse SC.

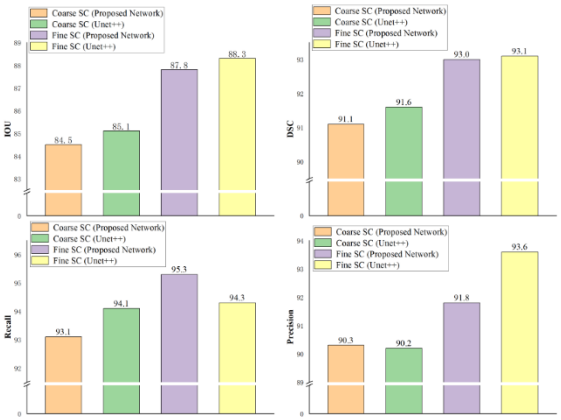

FIGURE S8. Comparison of quantitative metrics for our segmentation network and Unet++ for center A.

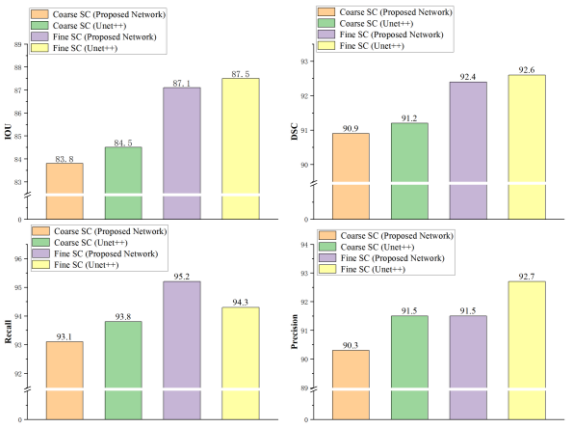

FIGURE S9. Comparison of quantitative metrics for our segmentation network and Unet++ for center B.

Classification results

Comparison of classification results of MC-DSCN, ResNet50 and VGG16

Each column from left to right in Figure S11 are the ROC curves of the classification component of proposed MS-DSCN, ResNet50 and VGG16 for different input modes (T2w with mask, ADC with mask, mp-MRI with mask, respectively). It can be noted from the figures in the first line that for both T2w and ADC images, the AUC with coarse masks is always higher than the AUC without coarse masks, and the AUC with coarse masks for MS-DSCN is the highest. And it can be noted from the figures in the second line that the AUC of mp-MRI data with coarse masks and co-training block is the highest. Overall, for all modes, the AUC of VGG16 is the lowest, and the AUC of MS-DSCN and ResNet50 are similar. However, the number of network parameters of MS-DSCN (2.15M) is far smaller than ResNet50 (47.07M) and VGG16 (65.06M).

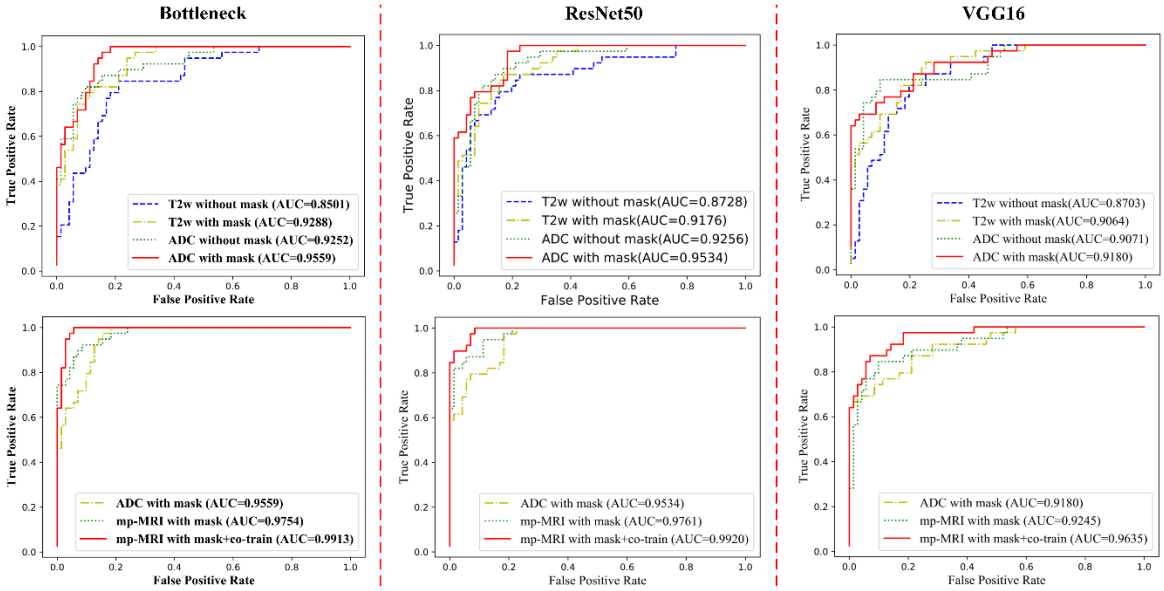

FIGURE S10. From top to bottom and from left to right are the Roc curves of the classification component of proposed MS-DSCN, ResNet50 and VGG16 for different input modes, respectively.

Evaluation and choice of hyper-parameters for hybrid loss in MC-DSCN

Evaluation and choice of hyper-parameters for hybrid loss during coarse segmentation

To evaluate loss function of the segmentation network, we trained our model using various loss functions, including generalized Dice loss (GDL loss), weighted cross-entropy (WCE loss), hybrid loss that combined GDL and WCE loss, and herein proposed loss (GDL loss + WCE loss + rank loss). Table S1 summarizes the segmentation results using different loss functions for the coarse segmentation network. GDL loss is region-based loss to profile segmentation results at the image level. WCE loss is a distribution-based loss that attempts to constrain segmentation results at the pixel level. Combination of these two losses outperforms each one of them applied separately, namely, the IOU of the network with the hybrid loss is 83.60%, and the IOU only with GDL loss is 82.12% ($p < 0.05$). Our proposed loss also contains one boundary-based loss to ensure the accuracy of the boundary pixels. Compared to the hybrid loss with only GDL and WCE, our proposed loss improves the segmentation results, demonstrated by the IOU increase from 83.60% to 84.14% ($p < 0.05$).

We also investigated hyper-parameters of the weighting factor $\lambda_1, \lambda_2, \lambda_3$ in equation 1 that controls the contribution of $Loss_{GDL}$, $Loss_{WCE}$ and $Loss_{rank}$. Table S2 summarized the metrics of the coarse segmentation network while increasing λ_3 ($Loss_{rank}$) from 0.00 to 0.10. Based on these results, we chose to set $\lambda_1 = 1.0, \lambda_2 = 1.0, \lambda_3 = 0.1$ as the default weighting factor for our segmentation hybrid loss for the experiments throughout the study.

TABLE S1. Comparison of the results of coarse segmentation network with different loss functions

Loss Function	DSC	IOU	Recall	Precision
GDL loss	89.50	82.12	91.37	89.57
WCE loss	89.58	82.20	92.63	87.89
GDL loss + WCE loss	90.44	83.60	93.13	89.50
Our proposed loss	**91.12**	**84.50**	**93.14**	**90.31**

TABLE S2. Quantitative metrics of the coarse segmentation network as a function of λ_3 ($Loss_{rank}$)

Hyper-parameters set	DSC	IOU	Recall	Precision
$\lambda_1 = 1.0, \lambda_2 = 1.0, \lambda_3 = 0.00$	90.44	83.60	93.14	89.50
$\lambda_1 = 1.0, \lambda_2 = 1.0, \lambda_3 = 0.05$	90.76	84.09	93.71	89.34
$\lambda_1 = \mathbf{1.0}, \lambda_2 = \mathbf{1.0}, \lambda_3 = \mathbf{0.10}$	**91.12**	**84.50**	93.14	**90.31**
$\lambda_1 = 1.0, \lambda_2 = 1.0, \lambda_3 = 0.15$	90.58	83.84	93.24	89.59
$\lambda_1 = 1.0, \lambda_2 = 1.0, \lambda_3 = 0.20$	90.50	83.63	93.03	89.56
$\lambda_1 = 1.0, \lambda_2 = 0.5, \lambda_3 = 0.10$	90.39	83.81	93.47	89.31
$\lambda_1 = 0.5, \lambda_2 = 1.0, \lambda_3 = 0.10$	90.41	83.45	**95.14**	87.12

Evaluation and choice of hyper-parameters of hybrid loss for classification

As mentioned in section (Network optimizing and training) of the manuscript, our classification network utilizes a hybrid loss function that not only contains the cross-entropy loss of T2w and ADC images, but also an inconsistency loss to ensure consistency of CRMs derived from ADC and T2w images. Here, we investigate the optimal choice of hyper-parameters (weighting factors $\lambda_1, \lambda_2, \lambda_3$ in equation 7) that controls the contribution of the classification loss functions, and the normalized inconsistency loss function. Table S3 summarizes the metrics of the classification network as a function of λ_3 (normalized inconsistency loss) in the range from 0.00 to 1. Based on these results, we set $\lambda_1 = 1.0, \lambda_2 = 1.0, \lambda_3 = 0.25$ as the default weighting factor for our classification hybrid loss for the experiments throughout the study.

TABLE S3. Quantitative metrics of the classification network as a function of λ_3 (inconsistency loss)

Hyper-parameter set	Precision	Sensitivity	Specificity	AUC
$\lambda_1 = 1.0, \lambda_2 = 1.0, \lambda_3 = 0.00$	0.8085	0.9513	0.8718	0.9753
$\lambda_1 = 1.0, \lambda_2 = 1.0, \lambda_3 = 0.15$	0.8586	0.9711	0.9083	0.9887
$\lambda_1 = \mathbf{1.0}, \lambda_2 = \mathbf{1.0}, \lambda_3 = \mathbf{0.25}$	**0.8963**	0.9727	**0.9356**	**0.9913**
$\lambda_1 = 1.0, \lambda_2 = 1.0, \lambda_3 = 0.50$	0.8835	0.9519	0.9260	0.9895
$\lambda_1 = 1.0, \lambda_2 = 1.0, \lambda_3 = 0.75$	0.8546	0.9633	0.8953	0.9872
$\lambda_1 = 1.0, \lambda_2 = 1.0, \lambda_3 = 1.00$	0.8294	**0.9927**	0.8649	0.9845
$\lambda_1 = 0.5, \lambda_2 = 1.0, \lambda_3 = 0.25$	0.8043	0.9487	0.8732	0.9776
$\lambda_1 = 1.0, \lambda_2 = 0.5, \lambda_3 = 0.25$	0.8292	0.8717	0.9014	0.9750

TABLE S4. The comparison results of coarse segmentation network with different loss functions

Loss Function	DSC	IOU	Recall	Precision
Dice loss	89.34	81.72	93.64	87.45
Focal_tversky loss	88.54	81.07	93.09	86.62
Focal loss	89.91	82.59	**96.08**	85.62
IOU loss	89.37	82.31	92.16	88.97
BCE + Dice loss	90.23	83.07	92.74	89.23
Our proposed loss	**91.12**	**84.50**	93.14	**90.31**

The ablation experiment

The ablation experiment for classification with T1w images

To evaluate T1 weighted image for classification, we performed the ablation experiment with or without T1 weighted image for classification, and the results are shown in Figure S12. Performance of classification is not improved with T1 weighted image (98.6% with T1w, T2w and ADC, and 99.1% with T2w and ADC).

We think the reason for not increasing AUC with additional T1w may be due to T1w and T2w providing similar anatomical information. As reported in Prostate Imaging Reporting and Data System Version 2.0 (PI-RADS 2.0) (a standardized PCa reporting system published in 2015 by the American College of Radiology), T2W and ADC from DWI are the two most recommended dominant sequences for Pca classification. Specifically, it pointed out that T2w which reflects the anatomy of a prostate is useful for delineating suspicious lesions because of their shorter "T2 relaxation time" compared with normal glandular tissue. And DWI which reflects the degree of water diffusion has low apparent diffusion coefficient (ADC) values in cancerous regions due to the tightly packed cells. Moreover, several studies have shown significant increases in both sensitivity and specificity by combining both ADC and T2w[3-5]. Thus, in this work, we also only utilized the ADC and T2w images.

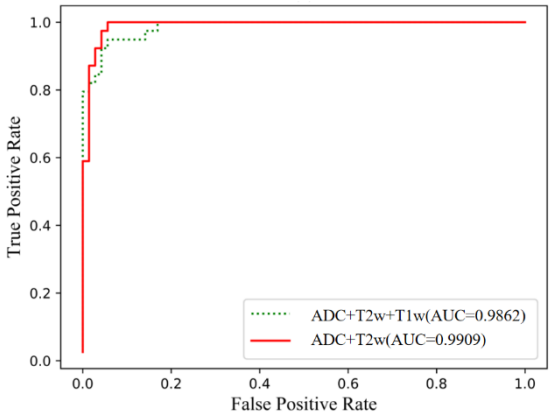

FIGURE S11. The ROC curves of the classification component for different input modes (ADC+T2w+T1w, ADC+T2w)

The ablation experiment for classification with only T2w images or ADC images

We compared the CRMobtained with only T2w or ADC in Figure S13. The CRM obtained with T2w is different from CRM obtained with ADC. We think that the CNN of classification sees' not only PCa-relevant patterns but also irrelevant visual patterns when distinguishing slices containing PCa from the normal ones. In addition, the irrelevant patterns are different for CNN ADC and CNN T2w, and in turn result in different CRMs[5]. To address this problem, we enforce the CNN models of ADC and T2w to generate consistent prediction labels and similar CRMs. As the irrelevant visual patterns learned from CNN ADC are different from those learned from CNN T2w while PCa-relevant patterns learned from both models are similar, the enforcement can greatly reduce the amount of irrelevant visual patterns without losing PCa-relevant patterns.

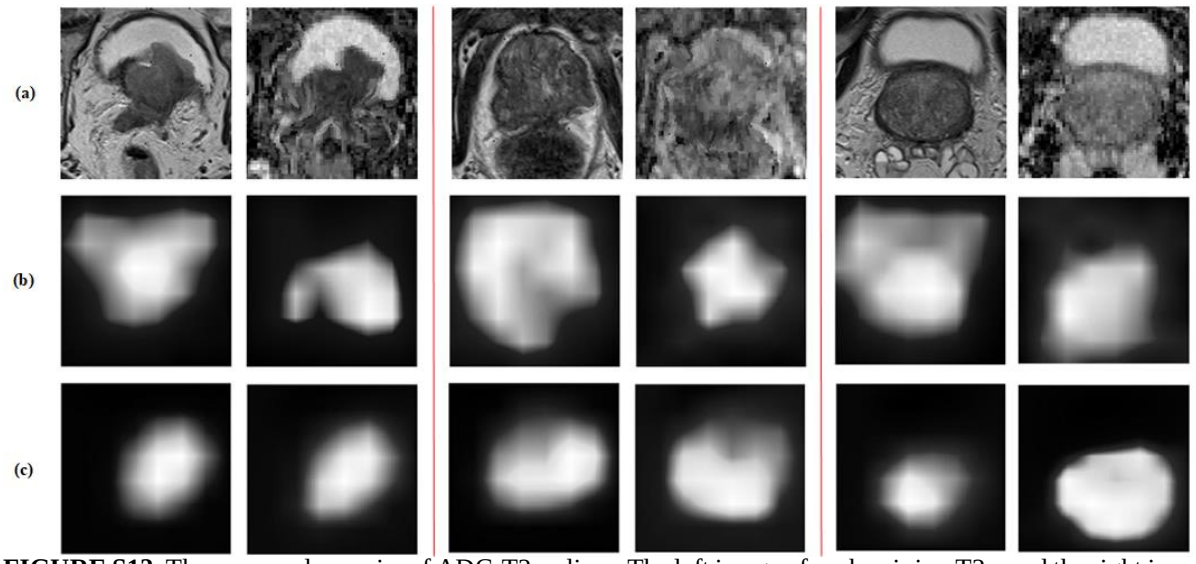

FIGURE S12. Three exemplary pairs of ADC-T2w slices. The left image of each pair is a T2w and the right image of each pair is an ADC. (b) The cancer response maps of T2w and ADC based on single-modality CNN models. Clearly, the CRM of T2w is inconsistent with that of ADC. (c) The cancer response maps for T2w and ADC based

on our co-trained CNN model. T2w and ADC together have much more consistent CRMs and the responses at cancerous regions are enhanced for both ADC and T2w compared to those based on a single-modality CNN.

References

1. R. Girshick, J. Donahue, T. Darrell and J. Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 580-587.
2. Weiwei Du, Shiyang Wang, A. Oto and Yahui Peng, "Graph-based prostate extraction in T2-weighted images for prostate cancer detection," 2015 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), 2015, pp. 1225-1229.
3. Fehr D, Veeraraghavan H, Wibmer A, Gondo T, Matsumoto K, Vargas HA, Sala E, Hricak H, Deasy JO. Automatic classification of prostate cancer Gleason scores from multiparametric magnetic resonance images. Proc Natl Acad Sci U S A. 2015 Nov 17;112(46):E6265-73.
4. Peng Y, Jiang Y, Yang C, Brown JB, Antic T, Sethi I, Schmid-Tannwald C, Giger ML, Eggener SE, Oto A. Quantitative analysis of multiparametric prostate MR images: differentiation between prostate cancer and normal tissue and correlation with Gleason score--a computer-aided diagnosis development study. Radiology. 2013 Jun;267(3):787-96.
5. Yang X, Liu C, Wang Z, Yang J, Min HL, Wang L, Cheng KT. Co-trained convolutional neural networks for automated detection of prostate cancer in multi-parametric MRI. Med Image Anal. 2017 Dec;42:212-227.