

Article

Interactive Cascaded Network for Prostate Cancer Segmentation from Multimodality MRI with Automated Quality Assessment

Weixuan Kou ¹, Cristian Rey ², Harry Marshall ³ and Bernard Chiu ^{4,*}¹ Department of Electrical Engineering, City University of Hong Kong, Hong Kong; weixuakou2-c@my.cityu.edu.hk² Schulich School of Medicine & Dentistry, Western University, London, ON N6A 5C1, Canada; cristian.rey@lhsc.on.ca³ Department of Radiology, Vanderbilt University Medical Center, Nashville, TN 37232, USA; harry.marshall@vumc.org⁴ Department of Physics & Computer Science, Wilfrid Laurier University, Waterloo, ON N2L 3C5, Canada

* Correspondence: bchiu@wlu.ca

Abstract: The accurate segmentation of prostate cancer (PCa) from multiparametric MRI is crucial in clinical practice for guiding biopsy and treatment planning. Existing automated methods often lack the necessary accuracy and robustness in localizing PCa, whereas interactive segmentation methods, although more accurate, require user intervention on each input image, thereby limiting the cost-effectiveness of the segmentation workflow. Our innovative framework addresses the limitations of current methods by combining a coarse segmentation network, a rejection network, and an interactive deep network known as Segment Anything Model (SAM). The coarse segmentation network automatically generates initial segmentation results, which are evaluated by the rejection network to estimate their quality. Low-quality results are flagged for user interaction, with the user providing a region of interest (ROI) enclosing the lesions, whereas for high-quality results, ROIs were cropped from the automatic segmentation. Both manually and automatically defined ROIs are fed into SAM to produce the final fine segmentation. This approach significantly reduces the annotation burden and achieves substantial improvements by flagging approximately 20% of the images with the lowest quality scores for manual annotation. With only half of the images manually annotated, the final segmentation accuracy is statistically indistinguishable from that achieved using full manual annotation. Although this paper focuses on prostate lesion segmentation from multimodality MRI, the framework can be adapted to other medical image segmentation applications to improve segmentation efficiency while maintaining high accuracy standards.

Keywords: prostate lesion segmentation; automatic quality assessment; interactive segmentation



Citation: Kou, W.; Rey, C.; Marshall, H.; Chiu, B. Interactive Cascaded Network for Prostate Cancer Segmentation from Multimodality MRI with Automated Quality Assessment. *Bioengineering* **2024**, *11*, 796. <https://doi.org/10.3390/bioengineering11080796>

Academic Editor: Giuseppe Baselli

Received: 22 June 2024

Revised: 17 July 2024

Accepted: 30 July 2024

Published: 6 August 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Prostate cancer (PCa) is the most common type of cancer and the second leading cause of cancer-related deaths among men in the United States, accounting for approximately 29% of new cancer cases and 11% of all cancer deaths in 2023 [1]. The five-year survival rate of PCa exceeds 97% with early diagnosis and treatment [1], highlighting the critical need for efficient and accurate PCa detection methods in clinical practice. Conventional screening methods, such as digital rectal examination (DRE) and blood tests for prostate-specific antigen (PSA), are highly sensitive but have low specificity [2], which often leads to overtreatment. After these initial screenings, the standard diagnostic procedure is a systematic transrectal ultrasound (TRUS)-guided biopsy. However, this method has several drawbacks, primarily its inability to precisely visualize and target prostate lesions. This limitation often results in missed cancers or unnecessary prostate biopsies, potentially causing unnecessary pain to patients [2].

Multiparametric magnetic resonance imaging (mpMRI) has shown superior performance in distinguishing clinically significant PCa (csPCa) from benign conditions, providing more precise PCa localization than ultrasound imaging [3,4]. Consequently, mpMRI has gained increasing attention in the detection and diagnosis of PCa. In order to standardize the use of mpMRI on prostate disease diagnosis, Prostate Imaging—Reporting and Data System (PI-RADS) [5–8] was developed as a consensus guideline, which recommends the combination of T2-weighted (T2W) MRI with functional MRI sequences, such as diffusion-weighted imaging (DWI) and dynamic contrast-enhanced (DCE) MRI, to stage and delineate PCa. Additionally, the apparent diffusion coefficient (ADC) sequence is commonly used in clinical practice and is derived from the DWI sequences of different b-values [5].

Although mpMRI is a powerful non-invasive technique for detecting PCa, and mpMRI-guided biopsies can significantly improve diagnosis accuracy while reducing unnecessary pain to patients, the analysis of mpMRI is time-consuming and subject to significant inter-observer variability due to differences in radiologist expertise. Therefore, computer-aided MRI PCa segmentation methods are highly demanded in clinical practice. Recent advancements in deep learning, particularly convolutional neural network (CNN)-based methods [9,10], have shown superior performance in medical image segmentation. However, current automated methods [9–19] have not attained sufficient accuracy and robustness in all cases. This limitation primarily stems from inherent challenges in prostate mpMR images, including low tissue contrast, irregular shapes, ambiguous boundaries, and large variations across patients [20].

In order to address these issues, we propose an interactive cascaded framework for PCa segmentation. The central hypothesis of our approach is that human intervention in automated segmentation approaches will provide a leap in lesion identification and segmentation accuracy. However, the need for a human observer to review all axial images and provide annotation would have been time-consuming. Our strategy for improving efficiency involves two crucial steps. First, we train a network to automate region of interest (ROI) detection, providing useful information to alleviate expert manual labor. Second, we automate the flagging of “inaccurate” ROI to avoid the need for users to manually review all results. The expert user can then focus on editing only the ROI for these inaccurate cases, thereby improving the cost-effectiveness of the review workflow. However, because the automatic evaluation of the quality of an ROI is a nontrivial problem itself, we propose a regression network we call rejection network that rates the automatically generated lesion segmentation result and flags low-quality detection results for user intervention. Users would identify a bounding box enclosing the lesion, and this bounding box would then assist another network to generate the final lesion boundary. Notably, this approach involves the delicate balance between the observer’s effort in editing the ROI and the final segmentation performance. That is, increasing the number of images edited by the expert observer will improve segmentation performance and will decrease cost-effectiveness. We conducted extensive experiments to evaluate the trade-off between accuracy and cost-effectiveness. The results we present allow users to choose a suitable cutoff point to balance the amount of user interaction and the segmentation accuracy desired for an application.

In this study, we used T2W, ADC, and DWI with a high b-value (DWI_{hb}) for PCa segmentation, opting to exclude the DCE sequence due to its longer acquisition time and the risk of allergic reactions to intravenous contrast in clinical practice [5]. The inclusion of the DCE sequence has been reported to achieve similar diagnostic accuracy as cases without DCE [8]. In PI-RADS version 2.1, the role of DCE is limited to that of a tiebreaker between a PI-RADS score of 3 and 4 in the peripheral zone [8]. We evaluated the proposed method on two publicly available prostate datasets. The experimental results demonstrate that the proposed method outperforms existing state-of-the-art automated PCa segmentation methods and attains competitive performance with significantly fewer human interventions compared to state-of-the-art fully interactive segmentation methods.

2. Related Works

2.1. Deep Learning for Prostate Cancer Segmentation

Deep learning-based automated segmentation methods [9–19] have shown promise in effectively detecting and staging PCa from mpMRI by automatically learning abundant semantic features. Wang et al. [11] proposed a dual-path CNN framework that separately extracts T2W and ADC features and then fuses them to generate the final results. Instead of solely fusing features before the final classifier, Zhang et al. [16] developed a dual-path self-attention distillation network to extract attention maps for T2W and ADC features in each feature layer of the encoder and then fuse them within both the current and adjacent layers. They further used the Kullback-Leibler divergence loss to enforce agreement in T2W and ADC feature extraction at each feature level. Jiang et al. [19] proposed a three-stream fusion U-Net to separately extract T2W, ADC, and DWI features and then fuse these features in each layer of the encoder. An adaptive weight was introduced for each stream, enabling the network to automatically highlight the most relevant stream.

The above methods [11,16,19] process each MRI sequence individually and design sophisticated modules to fuse information from different streams. Isensee et al. [10] proposed nnU-Net and demonstrated that concatenating various MRI sequences along the channel dimension in the input image, combined with appropriate preprocessing, can achieve state-of-the-art performance across multiple tasks. nnU-Net is more computationally efficient than processing each sequence separately. Mehravand et al. [17] conducted experiments to investigate PCa segmentation performance with an automated cascaded segmentation framework on biparametric MRI. They reported a mean Dice similarity coefficient (DSC) of 0.359 and identified a high number of false positives in predictions as a significant impediment to model performance. In order to address this, they trained an independent network to filter false positives in predictions, but the improvements were limited [17].

Reducing false positives and false negatives in predictions is the key to improving PCa segmentation performance. Integrating human expertise and automated approaches is a promising direction for achieving this. Interactive segmentation algorithms incorporate human expertise into CNN segmentation models to highlight specific regions, thereby achieving robustness and superior performance on various tasks of medical image analysis [21]. Kirillov et al. [22] proposed the Segment Anything Model (SAM) to transform the relative coordinates of user interactions into sinusoidal-like embeddings, integrating them into the decoder to enhance focus on specific regions of interest in the deep-layer features. By incorporating human interactions, SAM outperforms state-of-the-art automated methods in various tasks, demonstrating the effectiveness of human knowledge in improving the performance of CNN-based methods. However, SAM requires users to provide interactions with each input image, which limits its cost-effectiveness in medical imaging applications, where hundreds of images are required to be segmented for clinical analyses. A potential approach to improve the cost-effectiveness of SAM is to automate ROI detection but allow users to edit if the ROI is not accurate. This strategy would allow users to focus only on cases where automated segmentation methods [9–19,23] underperform. As illustrated in Figure 1, the automated segmentation method performs well in Figure 1d,e but fails in Figure 1a–c. If bounding boxes were automatically extracted from the results in Figure 1d,e, they could replace human inputs in SAM for efficient and accurate segmentation. On the other hand, the low-quality ROI generated for Figure 1a–c would not result in accurate segmentation by SAM. The key is in the development of an assessment workflow for automatically generated ROI that would allow for the accurate flagging of low-quality ROI for user intervention.

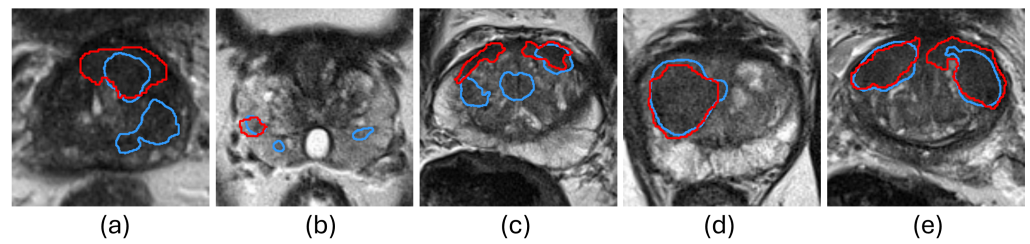


Figure 1. Examples of PCa segmentation results from U-Net [9]. The red and blue contours represent ground truth and algorithm segmentation, respectively. (a–c) show cases with false positives and false negatives generated by the automated segmentation method, and (d,e) show cases where the automated segmentation method demonstrates superior performance.

2.2. Automated Quality Assessment of Segmentation Results

Although various algorithms [9,10,17,22] have been proposed to generate segmentation results for medical images, the evaluation of segmentation quality is typically carried out manually by experts. Manually reviewing the quality of each result is impractical for the clinical implementation of these algorithms. Therefore, the automated quality assessment of segmentation results generated by CNN-based algorithms is a crucial step in streamlining user-friendly computer-aided diagnosis [24]. Jiang et al. [25] proposed an IoU-Net to regress the intersection-over-union (IoU) of the bounding box-level metrics of the predicted results. Huang et al. [26] advanced this approach by proposing the Mask Scoring R-CNN, which utilizes instance-level IoU to quantify the quality of region proposals in Mask R-CNN [27]. Eidex et al. [28] integrated a mask scoring head of [26] into their cascaded CNN to improve the performance of coarse-to-fine segmentation. Although the predicted IoU scores can approximate the actual segmentation quality, as reported in [26], these methods [25,26,28] only integrate IoU scores into the model to adaptively refine the low-quality results without investigating the potential of integrating quality scores into a human-in-the-loop segmentation. Zhou et al. [24] used a similar mask scoring head as [26] to rank the segmentation quality of axial images within a 3D volumetric image to allow user input for iterative segmentation enhancement. As their goal was just to rank the segmentation quality of axial images within a 3D volumetric image, they did not consider the problem of providing a consistent quality score across patients. A consistent assessment across patients is critical in a medical image segmentation workflow, and this is what we aim to establish in this paper.

3. Method

3.1. Overall Workflow

The overall workflow of our proposed interactive cascaded network is illustrated in Figure 2, consisting of a coarse segmentation U-Net [9], a rejection network, and a SAM [22] to generate the final segmentation. Let $x_i \in \mathbb{R}^{3 \times H \times W}$, $i \in \{1, 2, \dots, N\}$ be an input image block with an axial spatial size of $H \times W$, and the three channels represent the T2W, ADC, and DWI_{hb} image modalities. All entities with the subscript i in the description below are associated with the i th image, and N represents the total number of images. x_i is first fed to the coarse segmentation U-Net, generating the prediction $p_i^c \in \mathbb{R}^{2 \times H \times W}$ as follows:

$$p_i^c = \theta_C(x_i), \quad (1)$$

where θ_C represents the coarse segmentation U-Net. The two channels in p_i^c provide background and foreground segmentation results. We denote f_i^{final} as the features extracted by the model before the final classifier stage of θ_C . The concatenation of f_i^{final} and p_i^c serves as the input of the rejection network, which can be represented by

$$s_i = \theta_R(cat(f_i^{final}, p_i^c)), \quad (2)$$

where θ_R represents the rejection network, and cat represents the concatenation operation. s_i is the segmentation quality score generated by θ_R . The Dice similarity coefficient (DSC)

between the prediction p_i^c and the segmentation ground truth y_i is used as the target quality score to be estimated in our study.

Segmentation results with low s_i are flagged for the users to input a bounding box enclosing the lesion; for segmentation results with high s_i , bounding boxes are automatically cropped from the boundary segmented by the coarse segmentation U-Net. The manually and automatically generated bounding boxes are fed to the SAM for further refinement, resulting in the final segmentation results p_i^f :

$$p_i^f = \theta_{\mathcal{F}}(x_i, \text{bbox}(p_i^c | s_i)), \quad (3)$$

where $\theta_{\mathcal{F}}$ indicates the fine segmentation SAM. $\text{bbox}(p_i^c | s_i)$ represents the automatically and manually generated upper left and lower right coordinates of the bounding box derived from p_i^c based on s_i .

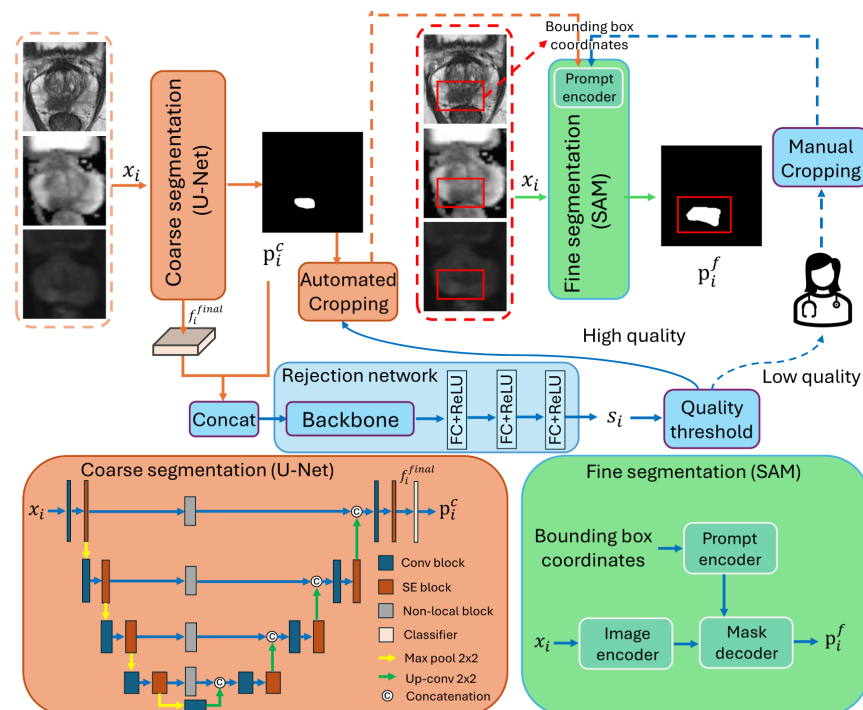


Figure 2. The workflow of the proposed interactive cascaded network.

3.2. Network Structure

3.2.1. Coarse Segmentation Network

The coarse segmentation network has a similar structure to U-Net [9] but with the Squeeze-and-Excitation (SE) [29] and non-local blocks [30] integrated to enhance the feature representations, as shown in Figure 2. The encoder path comprises five convolutional blocks with four pooling operations, as in [9], resulting in an output stride of 16. The SE blocks perform global average pooling followed by two fully connected layers to adjust channel-wise feature scaling dynamically, whereas non-local blocks enhance the spatial attention of features in skip connections by capturing long-range dependencies in an image.

3.2.2. Rejection Network

The rejection network was proposed to estimate the quality of the boundaries segmented by the coarse segmentation network so that low-quality segmentation results can be flagged for human intervention. EfficientNet-B5 [31] was used as the backbone to balance performance and computational efficiency [31]. Three fully connected (FC) layers with ReLU activations from the regression head were used to estimate the quality score of each image. We set the outputs of the first two fully connected layers of the regression head

to 1024 and 512, respectively. The output of the final fully connected layers is set to 1 to output the estimated quality score.

Images with a score lower than the threshold t are flagged for user intervention. Users can input a bounding box enclosing the lesion. For images with a score higher than the threshold, the bounding boxes are automatically determined based on the boundary segmented by the coarse segmentation U-Net. All bounding boxes are expanded by 40%. The expansion of the automatically determined boundary boxes provides a margin of error for the coarse segmentation algorithm. For manual identification, the expansion reduces the dependence on the accuracy in the bounding box selection, which may vary across users with different levels of expertise in prostate MRI.

3.2.3. Fine Segmentation Network

In our framework, both manually and automatically cropped bounding boxes can be treated as hints to guide the fine segmentation network in highlighting the specific regions. SAM [22] was used as the fine segmentation network to incorporate the generated bounding box information and enhance model performance. SAM consists of an image encoder, a prompt encoder, and a mask decoder, as detailed in [22]. We use the ViT-base model [22] as the image encoder to balance segmentation performance and computational efficiency, given that larger ViT models offer only marginal improvements in accuracy while significantly increasing computational cost [22,32]. The image encoder takes input images as inputs, while the upper left and lower right coordinates of the generated bounding box serve as inputs for the prompt encoder.

3.3. Model Training

The three networks involved in the proposed workflow were trained independently. The coarse segmentation U-Net is first trained with Dice loss [19], denoted by L_c :

$$L_c = 1 - \frac{1}{N} \sum_{i=1}^N \frac{2 \sum_{j=1}^{H \times W} p_{i,j}^c y_{i,j}}{\sum_{j=1}^{H \times W} (p_{i,j}^c + y_{i,j})}, \quad (4)$$

where both y_i and p_i^c have the same size as the input image x_i , which has a height of H and a width of W . y_i is a binary segmentation map indicating whether each pixel of the input image is within a manually segmented lesion, whereas p_i^c is the probability map generated by U-Net for each pixel being inside the lesion. $p_{i,j}^c$ and $y_{i,j}$ represents the j^{th} pixel in p_i^c and y_i , respectively. N represents the number of input images.

Once optimized, the parameters of the coarse segmentation U-Net were frozen, and the mean square error (MSE) loss was used to train the rejection network, denoted by L_r :

$$L_r = \frac{1}{N} \sum_{i=1}^N (s_i - d_i)^2, \quad (5)$$

where d_i is the DSC calculated between p_i^c and y_i , which is considered the target quality score, and s_i is the quality score of p_i^c predicted by the rejection network. N represents the number of input images.

SAM was initialized with the publicly available pretrained ViT-base model [22]. As the image encoder and prompt encoder were intensively trained and shown to be generalizable [22], we follow [32] in not further training them. With these encoders frozen, the decoder was fine-tuned with L_f , which is expressed below and is detailed in [32]:

$$L_f = -\frac{1}{N} \sum_{i=1}^N \sum_{j=1}^{H \times W} [y_{i,j} \log(p_{i,j}^f) + (1 - y_{i,j}) \log(1 - p_{i,j}^f)] + \left(1 - \frac{1}{N} \sum_{i=1}^N \frac{2 \sum_{j=1}^{H \times W} p_{i,j}^f y_{i,j}}{\sum_{j=1}^{H \times W} (p_{i,j}^f + y_{i,j})} \right), \quad (6)$$

where both y_i and p_i^f have the same size as the input image x_i , which has a height of H and a width of W . p_i^f is the prediction of SAM, and y_i is the corresponding ground truth. $p_{i,j}^f$

and $y_{i,j}$ represents the j^{th} pixel in p_i^f and y_i , respectively. N represents the number of input images. We note that the first term of L_f is the cross-entropy loss [9], and the second term is the Dice loss [19].

4. Experimental Setup

4.1. Data

We evaluated our method on the following two publicly available datasets. The annotations for Prostate158 are publicly available, while the segmentation of the prostate lesion in PROSTATEx2 was performed by us.

Prostate158: This dataset comprises 158 annotated prostate MR images acquired by 3-T Siemens scanners [33]. All contributions were approved by the institutional ethics committee, as stated in [33]. The segmentation annotations of prostate zones and cancer were manually created by two board-certified radiologists [33]. The ADC and DWI sequences were resampled to the orientation and voxel spacing of the T2W sequence [33]. The voxel sizes of the T2W images vary from 0.272 mm to 0.495 mm, with a thickness from 3 mm to 3.15 mm. The T2W, ADC, and DWI images were further resampled to the median voxel spacing of $0.4 \times 0.4 \times 3 \text{ mm}^3$. Following the data partition in [33], the dataset was split into training (119 cases), validation (20 cases), and testing (19 cases) subsets on a patient basis.

PROSTATEx2: The training set of PROSTATEx2 [23,34,35] was used in our study, consisting of 99 prostate MR images acquired by 3T Siemens scanners. All contributions were approved by the institutional ethics committee, as stated in [23,34,35]. The segmentation annotations of PCa were manually performed by a medical student and a board-certified radiologist. The ADC and DWI sequences were resampled to the orientation and voxel spacing of the T2W sequence, achieving a $0.5 \times 0.5 \text{ mm}$ in-plane resolution and 3.6 mm slice thickness. We randomly split the dataset on a patient basis into training (60 cases), validation (9 cases), and testing (30 cases) sets.

4.2. Evaluation Metrics

Two metrics were used to evaluate the segmentation performance. The Dice similarity coefficient (DSC) between the algorithm segmentation, P , and the corresponding ground truth, G , is defined by $DSC = \frac{2|P \cap G|}{|P| + |G|}$, where $|\cdot|$ denotes the area of the operand. The Hausdorff distance (HD) evaluates the maximum distance from a point from ∂P (or ∂G) to its nearest point in ∂G (or ∂P). The 95%HD defined below is a variant designed to reduce the effect of outliers:

$$95\%HD = [\{\min_{j \in \partial G} d(i, j) : \forall i \in \partial P\}, \{\min_{i \in \partial G} d(i, j) : \forall j \in \partial P\}]_{95th}. \quad (7)$$

where $d(i, j)$ denotes the distance between points i and j .

An R-squared value (R^2) was used to evaluate the performance of the rejection network:

$$R^2 = 1 - \frac{\sum_{i=1}^N (d_i - s_i)^2}{\sum_{i=1}^N (d_i - \bar{d})^2}, \quad (8)$$

where s_i is the prediction of the rejection network, d_i is the DSC ground truth, and $\bar{d}$ is the mean of d_i .

The lesion-level recall and precision defined in [19,36] were applied to evaluate the model performance in lesion localization. Each lesion was considered a 3D entity, and all metrics were calculated on a lesion basis. Each of the I ground truth lesion volumes $\{G_i\}_{i=1}^I$ is considered a true positive (TP) if the overlap with any J algorithm segmented lesion volumes $\{P_j\}_{j=1}^J$, single or multiple, is larger or equal to a threshold, τ , and a false negative (FN) otherwise (i.e., G_i is a TP if $\sum_{j=1}^J |G_i \cap P_j| / |G_i| \geq \tau$, or a FN otherwise). The lesion-level recall (or true positive rate, TPR) is defined as $TP / (TP + FN)$. For the calculation of precision (or positive predictive value, PPV), the definition of TP is based on the algorithm segmentations P_j . That is, P_j is a TP if $\sum_{i=1}^I |G_i \cap P_j| / |P_j| \geq \tau$, or it is false positive (FP) otherwise. The precision is defined to be $TP / (TP + FP)$. We used a low

threshold $\tau = 0.1$ due to substantial interobserver variability in localizing lesions. Lesions can be underestimated even with manual segmentation, typically requiring margins up to 10 mm to be added to boundaries for focal treatments [37,38]. The F_1 score was calculated to summarize the lesion-level precision and recall as $F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$.

4.3. Implementation Details

Our model was implemented with the PyTorch library [39] and executed on an NVIDIA RTX3090 GPU. Adam optimizer was used to optimize the coarse segmentation U-Net and the rejection network with a batch size of 16. The initial learning rate was set to 0.001 for the coarse segmentation U-Net. For the rejection network, the initial learning rate was set to 0.0001 for the backbone and 0.001 for the regression head. The learning rate was decreased according to polynomial decay with an exponent of 0.9. We followed [32] to use AdamW [40] optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$), with an initial learning rate of 0.0001 and a weight decay of 0.01 to fine-tune the decoder part of SAM. The batch size was 4.

The coarse segmentation U-Net was trained for 200 epochs. All input images for coarse segmentation U-Net were uniformly resampled to the size of 256×256 . Random horizontal flipping, rotation, and cropping were applied to augment the training set. Once optimized, the coarse-segmentation U-Net was frozen, and the rejection network was trained for 200 epochs. The optimized coarse-segmentation U-Net and rejection network were used in the inference stage.

The fine segmentation SAM was first initialized with the publicly available pretrained ViT-base model [22]. The inputs to SAM included the images and bounding boxes generated from segmentation ground truth. The bounding boxes were randomly extended from 0% to 60% and then augmented with a random perturbation of 0–20 pixels in the training stage. All input images for SAM were uniformly resampled to the size of 256×256 . The decoder part of SAM was fine-tuned for 150 epochs [32].

The model with the highest DSC or the lowest MSE on the validation set was used for evaluation on the test set. The intensity of T2W, ADC, and DWI_{hb} was separately scaled into $[0, 1]$ and normalized. The threshold, t , in the rejection network is a flexible parameter that can be adjusted based on the practical requirements and model performance of the coarse segmentation network. In our study, we set $t = 0.4$ to show the model performance with a low burden of human intervention and $t = 0.7$ for a high burden of human intervention. Section 5.2 details the influence of t on the model performance and burdens of human intervention for PCa segmentation.

5. Experimental Results

5.1. Comparison with State-of-the-Art Methods

The segmentation performance of the proposed interactive cascaded network was compared with the following state-of-the-art methods: nnU-Net [10], UNETR [41], MIDeepSeg [42], and SAM [22]. Specifically, nnU-Net and UNETR serve as benchmarks for automatic segmentation methods without human interventions. MIDeepSeg and SAM are fully interactive segmentation methods incorporating human interventions in each input image, serving as the upper bound of the interactive segmentation methods. In our study, we use bounding box information as the user input of SAM since it has demonstrated superior performance on medical images [32]. The manually identified bounding box inputs for SAM were simulated by expanding the bounding box enclosing manual segmentations by 40%. The same strategy was used to simulate the manual boundary box in our approach for images with coarse segmentation quality scores lower than the threshold t . Bounding boxes are not applicable in MIDeepSeg [42], which requires extreme points on the boundaries of target regions to generate exponential geodesic distance maps. Hence, we follow the implementation in [42] to use extreme points as user inputs and set the number of extreme points as 4. All results are calculated on a per-image basis. Paired t-tests [43] were used to evaluate the statistical significance of performance difference in terms of DSC and 95% HD.

5.1.1. Comparison with Automated Segmentation Methods

Table 1 shows the DSC and 95% HD of the proposed method and the state-of-the-art automated segmentation methods nnU-Net [10] and UNETR [41] on the Prostate158 and PROSTATEx2 datasets. Model performance with $t = 0.4$ presents scenarios where a low burden of human intervention is required, as detailed in Section 5.2. At $t = 0.4$, 16.3% of the images from the Prostate158 test set and 20.9% from the PROSTATEx2 test set were flagged by the rejection network for user ROI identification. With manual ROI identification on approximately 20% of the images, the segmentation performance was significantly higher than automated segmentation methods (DSC: $p = 1.8 \times 10^{-8}$ for nnU-Net and 3.7×10^{-7} for UNETR in Prostate158, and $p = 9.4 \times 10^{-14}$ for nnU-Net and 3.0×10^{-13} for UNETR in PROSTATEx2; 95% HD: $p = 7.6 \times 10^{-4}$ for nnU-Net and 4.3×10^{-9} for UNETR in Prostate158, and $p = 2.7 \times 10^{-4}$ for UNETR in PROSTATEx2).

Table 1. Quantitative comparison with different methods on Prostate158 and PROSTATEx2 datasets in terms of slice-level DSC (%) and 95% HD (mm), and lesion-level recall (true positive rate, TPR, %), precision (positive predictive value, PPV, %) and F₁ score (%). The best results are highlighted in bold.

Methods	Prostate158					PROSTATEx2				
	DSC	95% HD	TPR	PPV	F ₁	DSC	95% HD	TPR	PPV	F ₁
nnU-Net	49.5 ± 32.4	10.4 ± 8.5	70.8	71.0	70.9	43.7 ± 34.3	12.6 ± 12.7	64.5	63.6	64.0
UNETR	51.0 ± 26.1	14.3 ± 10.7	75.0	42.1	55.1	44.3 ± 26.5	15.9 ± 10.6	71.0	49.4	58.3
MIDeepSeg	80.5 ± 10.6	8.2 ± 3.5	83.3	86.3	84.8	79.2 ± 11.9	3.0 ± 2.2	93.5	89.7	91.6
SAM	84.1 ± 8.6	6.5 ± 4.5	91.7	93.5	92.6	81.9 ± 10.7	10.8 ± 6.7	93.5	90.9	92.2
ours (t = 0)	49.3 ± 28.4	14.0 ± 12.1	62.5	46.9	53.6	42.5 ± 33.2	16.7 ± 14.3	51.6	44.8	48.0
ours (t = 0.4)	73.4 ± 16.2	8.7 ± 7.5	83.3	72.9	77.8	69.3 ± 21.4	14.0 ± 7.9	90.3	70.1	78.9
ours (t = 0.7)	83.2 ± 8.9	6.5 ± 4.6	91.7	91.8	91.7	81.0 ± 11.6	11.2 ± 7.0	93.5	88.3	90.8

Figure 3 shows a qualitative comparison, demonstrating the superior performance of the proposed framework. Figure 3c is a representative prostate image where nnU-Net and UNETR generated FP and FN predictions. For this example, the ROI was automatically generated by the coarse segmentation U-Net.

5.1.2. Comparison with Fully Interactive Segmentation Methods

Table 1 shows the DSC and 95% HD of the proposed method compared with the state-of-the-art fully interactive segmentation methods MIDeepSeg [42] and SAM [22] on the Prostate158 and PROSTATEx2 datasets. The performance by MIDeepSeg [42] and SAM [22] serves as the upper bound to assess the performance of our proposed framework with t set to 0.7, where images with a predicted DSC of less than 0.7 were provided a manually identified ROI. At $t = 0.7$, 46.9% of the images from the Prostate158 test set and 51.4% from the PROSTATEx2 test set were automatically flagged by the rejection network for user interaction. In this threshold setting, the burden of human intervention for our method was half of that involved in MIDeepSeg and SAM, which require user interaction for all images. Table 1 shows that SAM has the best segmentation performance on both datasets. In comparison, our proposed framework attains competitive performance with no significant statistical differences (DSC: $p = 0.075$ in Prostate158 and 0.058 in PROSTATEx2, 95% HD: $p = 0.49$ in Prostate158 and 0.06 in PROSTATEx2), but with approximately half the human intervention burden. Figure 4 shows a qualitative comparison in five example images, demonstrating that, with ROI manually identified in half of the images in the test set, our algorithm has a segmentation performance similar to the two methods requiring interactions in all images.

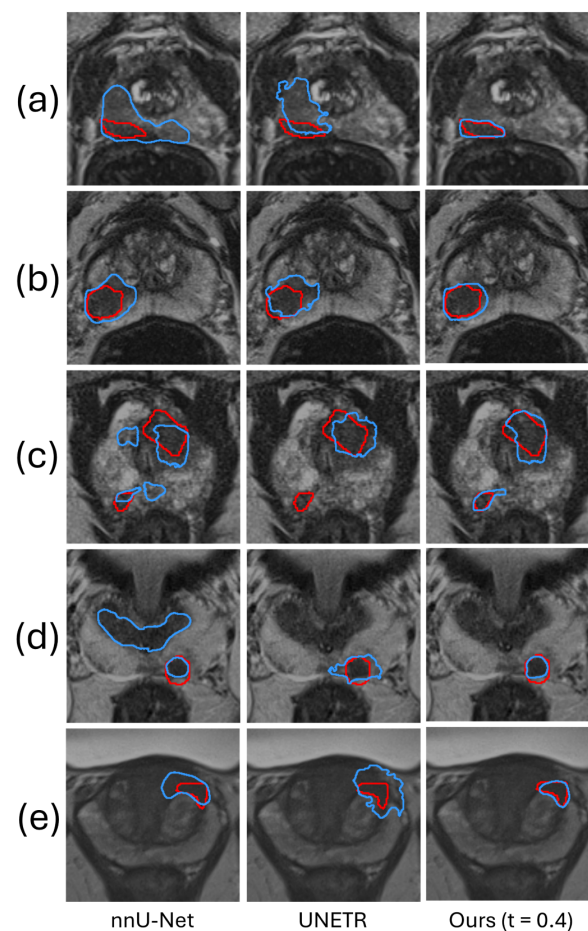


Figure 3. The segmentation results of our method are compared qualitatively with those produced by two automatic segmentation methods in five example images. The red and blue contours represent ground truth and algorithm segmentation, respectively. Prostate (a–c) are from the Prostate158 dataset, and Prostate (d,e) are from the PROSTATEx2 dataset.

5.1.3. Lesion-Level Analysis

Table 1 reports the TPR and PPV of different methods [10,22,41,42], as an assessment of lesion-level false positives (FPs) and false negatives (FNs). A higher recall indicates fewer FNs, while a higher precision indicates fewer FPs. The results in the first and second rows show that high FP is the major issue in automated segmentation methods, and incorporating human intervention substantially reduces both FP and FN (third and fourth rows). The performance of our method with $t = 0$ demonstrates that the FP and FN generated by the coarse segmentation can significantly affect the performance of the subsequent SAM. The TPR and PPV substantially improve with different levels of intervention, as shown in the results for the settings $t = 0.4$ and $t = 0.7$.

5.2. Analysis of Rejection Ratio

In our study, the R^2 value of the proposed rejection network is 0.75 in the Prostate158 and 0.71 in the PROSTATEx2. In order to quantify the workload required for manual ROI identification, we computed the percentage of the images requiring ROI identification, which we call the percentage rejection ratio. The rejection ratio is related to the threshold t , according to Figure 5a. It can be observed that the rejection ratio did not increase much when t was increased from 0.1 to 0.4. At this range, user involvement was less than 20%. For this reason, we used $t = 0.4$ to represent a low level of user involvement in Section 5.1. At $t = 0.7$, user involvement in ROI identification is approximately 50%. We used this level to represent a high level of user involvement in our comparisons in Section 5.1.

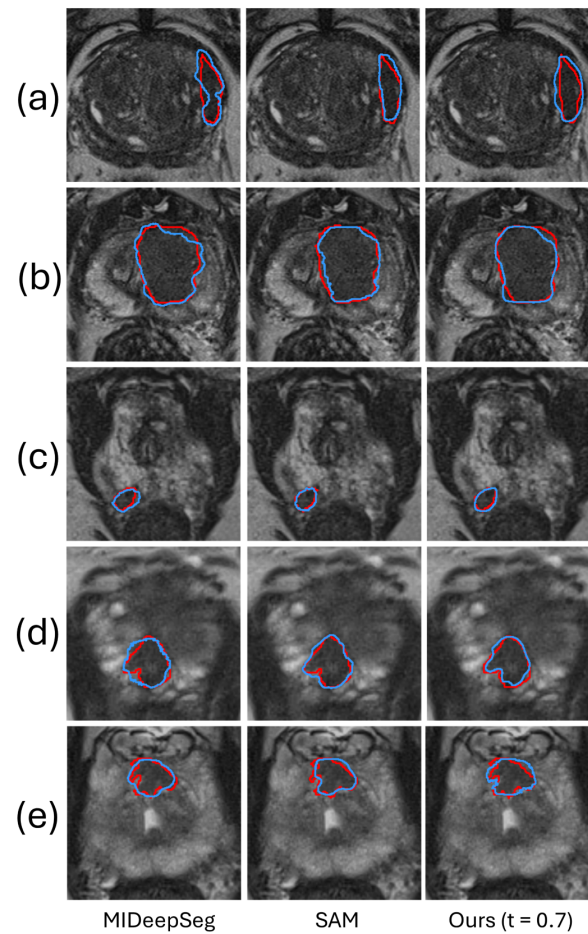


Figure 4. The segmentation results of our method are compared qualitatively with those produced by two fully interactive segmentation methods in five example images. The red and blue contours represent ground truth and algorithm segmentation, respectively. Prostate (a–c) are from the Prostate158 dataset, and Prostate (d,e) are from the PROSTATEx2 dataset.

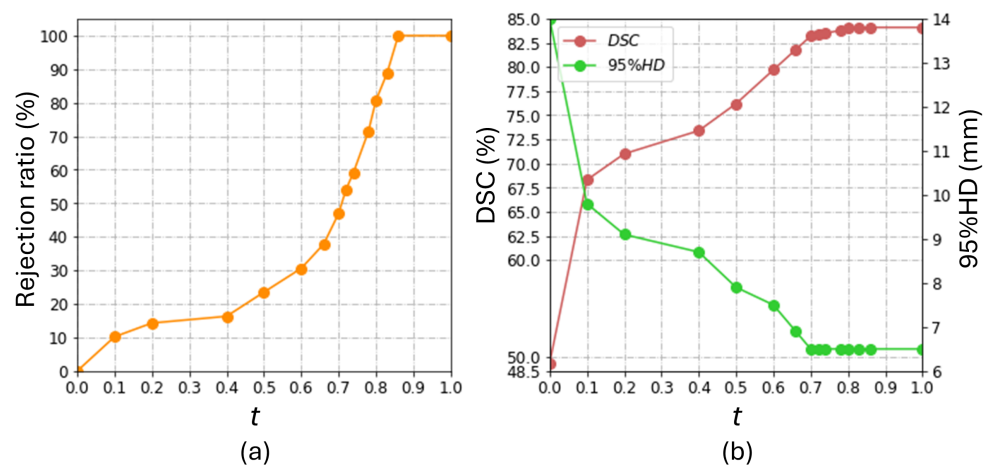


Figure 5. The influence of the burdens of human intervention on the model performance in the Prostate158 dataset. The rejection ratio quantifies the burden of human intervention. (a) shows the correlation between the threshold, t , and the rejection ratio in our study; (b) shows the effect of t on the overall model performance.

Figure 5b shows the overall performance at different levels of rejection ratio in the Prostate158 dataset. This figure shows that DSC and 95% HD enter a plateau with $t > 0.7$.

We attribute this observation to the superior performance of the coarse segmentation network on images with $t > 0.7$, where the automatically generated bounding boxes are comparable to human inputs. Meanwhile, Figure 5a shows that the rejection ratio (i.e., the workload involved in ROI identification) drastically increases with t exceeding 0.7. These observations suggest that it is not cost-effective to set t to be higher than 0.7.

The coarse segmentation U-Net attained a DSC of $47.8 \pm 26.7\%$ and a 95% HD of 13.1 ± 10.3 mm on the Prostate158 dataset. When the bounding box information was directly generated from the coarse segmentation results without the rejection network (i.e., $t = 0$), the DSC is $49.3 \pm 28\%$, and the 95% HD is 14.0 ± 12.1 mm. There was limited improvement in terms of DSC and even degradation in terms of 95% HD compared to the coarse segmentation results (DSC: $p = 9.3 \times 10^{-3}$; 95% HD: $p = 1.2 \times 10^{-3}$). In contrast, the segmentation performance significantly improved even with $t = 0.1$, which involved introducing human interactions into 10.2% of the images, as shown in Figure 5. The comparison between the performance at $t = 0$ and $t = 0.1$ (DSC: $p = 6.3 \times 10^{-6}$; 95% HD: $p = 4.1 \times 10^{-8}$) suggests that low-quality ROI substantially affects the segmentation accuracy from SAM. The segmentation performance at $t = 0.7$, with 46.9% of the images provided with manually annotated ROIs, was not significantly different from that at $t = 1$ that requires manual annotated ROI for all images (DSC: $p = 0.075$; 95% HD: $p = 0.49$). This is a key result that demonstrates the contribution of the proposed rejection network in improving the cost-effectiveness of the segmentation workflow.

5.3. Influence of Different Bounding Box Extension Ratio

We introduced the bounding box extension ratio to simulate scenarios where PCa ROIs may not be identified accurately due to the variability of the expertise across observers. Table 2 shows the influence of the bounding box extension ratio on model performance, with results obtained using $t = 0.4$. The DSC and 95% HD degraded slightly when the bounding box extension ratio increased from 10% to 40%, but the difference was not statistically significant (DSC: $p = 0.14$; 95% HD: $p = 0.47$). Additionally, the lesion-level metrics remain unchanged when the bounding box extension ratio is lower than 60%, indicating that no additional false positives or false negatives are generated, with the expansion increasing from 10% to 60%. Therefore, we set the bounding box extension ratio to 40% in our experiments to maintain flexibility and model performance.

Table 2. Ablation study investigating the influence of bounding box extension ratio on the model performance in the Prostate158 dataset in terms of slice-level DSC (%) and 95% HD (mm), and lesion-level recall (True positive rate, TPR, %), precision (Positive predictive value, PPV, %) and F₁ score (%).

Bounding Box Extension Ratio	DSC	95% HD	TPR	PPV	F ₁
10%	74.0 ± 16.0	8.5 ± 7.4	83.3	72.9	77.8
20%	73.9 ± 16.0	8.5 ± 7.5	83.3	72.9	77.8
30%	73.6 ± 16.2	8.7 ± 7.5	83.3	72.9	77.8
40%	73.4 ± 16.2	8.7 ± 7.5	83.3	72.9	77.8
50%	72.7 ± 16.4	8.8 ± 7.5	83.3	72.9	77.8
60%	71.8 ± 16.7	9.0 ± 7.8	83.3	72.9	77.8
80%	69.4 ± 17.2	9.4 ± 8.0	83.3	71.2	76.8

5.4. Training and Inference Time

We trained all methods listed in Section 5.1 on an NVIDIA RTX3090 GPU with 24 GB memory. Table 3 shows the total training time (h) and average inference time per image (ms) of different methods [10,22,41,42]. Notably, the training time of nnU-Net is significantly longer due to the default configurations for generating ensemble predictions. We trained only the decoder of SAM, as noted in Section 3.3, and as such, the training time reported includes the time involved in decoder training only. The training time of our proposed method is the sum of the training time for U-Net (1.8 h), the rejection network (6.5 h), and SAM (15.8 h). The user interpretation time is not included in the presented results,

which constitutes the majority of the time cost in actual implementation. Although the average inference time of our method is longer than that of MIDeepSeg [42] and SAM [22], the difference is negligible compared to the user interpretation time. The time for radiologists to interpret PCa from mpMRI varies depending on the complexity of the case and the experience of the radiologist. Studies reported that the average interpretation time per image ranges from 1 to 4 min, and it may be even longer in clinical practice [44,45]. The inference time of our proposed method is negligible compared to clinical interpretation time. A major advantage of our approach is that human interpretation is only needed for images with low-quality U-Net segmentation, thereby allowing an expert observer to focus only on problematic cases and expedite the review. In our study, the proposed framework attained competitive segmentation performance with no significant difference compared to MIDeepSeg and SAM by incorporating human annotation in half of the images.

Table 3. The training and inference time of different methods. The user interpretation time is not included in the inference time reported for MIDeepSeg [42], SAM [22], and our proposed method.

Methods	Training Time (h)	Inference Time (Ms/Image)
nnU-Net	101.5	91.7
UNETR	14.6	24.4
MIDeepSeg	18.2	20.3
SAM	15.8	140.5
ours	1.8 + 6.5 + 15.8	175.7

6. Discussion

We developed an interactive cascaded framework to enhance PCa segmentation performance from mpMRI and improve the efficiency of human-in-the-loop segmentation. Although networks, such as MIDeepSeg [42] and SAM [22], have been proposed to involve human interaction to improve the segmentation performance of automatic segmentation networks, the need for providing annotations in all input images limits the cost-effectiveness of these approaches. The proposed framework streamlines the interactive workflow by integrating (a) the coarse segmentation U-Net to automatically generate an initial ROI, (b) the rejection network that automatically estimates the quality of initial segmentation results and flags low-quality results for the manual identification of an ROI and (c) SAM that generated the final segmentation based on the ROI generated either manually or automatically. We conducted extensive experiments to show the contribution of the proposed rejection network and the trade-off between accuracy and cost-effectiveness. The proposed framework has demonstrated superior performance in terms of DSC and 95% HD in two publicly available datasets with manually identified ROI only for a fraction of the images.

The performance of the proposed framework benefits from the rejection network, which leverages the strengths of both automated and interactive segmentation methods. Previous studies [25,26,28] mainly focused on automatically refining low-quality results using estimated quality scores but did not integrate these scores into a human-in-the-loop segmentation. Our framework provides a method to incorporate quality score estimation into interactive segmentation, enhancing the efficiency of human-in-the-loop segmentation. The proposed rejection network improves the model performance in two ways. First, it rates and flags results where the coarse segmentation network underperforms for human intervention, thereby improving the final segmentation. Second, it enables users to focus only on low-quality results, increasing the efficiency of human-in-the-loop segmentation. At $t = 0.4$, with 16.3% of the images from Prostate158 and 20.9% from PROSTATEx2 provided with manual intervention, the DSC increased by 24.1% and the 95% HD by 5.3 mm compared to $t = 0$ (i.e., no user intervention) for the Prostate158 dataset (DSC: $p = 9.4 \times 10^{-8}$; 95% HD: $p = 1.0 \times 10^{-10}$). Similarly, there was a 26.8% increase in DSC and a 2.7 mm improvement in 95% HD for the PROSTATEx2 dataset at $t = 0.4$ compared with $t = 0$ (DSC: $p = 5.8 \times 10^{-15}$; 95% HD: $p = 3.9 \times 10^{-5}$). At $t = 0.7$, with 46.9% of the images from Prostate158 and 51.4% from PROSTATEx2 provided manual annotations, our

proposed framework attained a DSC and a 95% HD that were not significantly different from those generated by SAM with all images annotated (i.e., at $t = 1$) (DSC: $p = 0.075$ in Prostate158 and 0.058 in PROSTATEx2; 95% HD: $p = 0.49$ in Prostate158 and 0.06 in PROSTATEx2). Based on our experimental results, we would suggest using $t = 0.7$ to flag low-quality results for an optimal balance of segmentation performance and cost-effectiveness. In clinical practice, the selection of t depends on various factors, including the time users can afford, their level of expertise, and the required accuracy. A lower t corresponds to fewer interventions and lower segmentation performance, yet human interventions in a small number of cases can still significantly enhance model performance compared to automated methods. Users with a lower level of expertise could set a lower t to focus on cases for which the coarse segmentation network generates low-quality segmentation results and leave the cases with medium segmentation alone if he/she is not confident in improving the automatically generated ROI. The required accuracy of the task is the most important factor in deciding the degree of user involvement. t should be set higher if the task requires high accuracy. Additionally, users can choose to identify ROI at locations important for cancer diagnosis instead of identifying the ROI for all images with t smaller than the threshold.

Although our method has achieved high segmentation accuracy, our slice-based framework does not consider the continuity of adjacent axial slices. However, evidence suggests that 3D CNNs perform worse than 2D CNNs when slice thickness is substantially larger than in-plane voxel sizes due possibly to the limited continuity, such as in the prostate dataset in the decathlon challenge reported in [10]. Although an investigation in the current dataset is warranted, this issue is potentially present in our dataset due to similar thicknesses in mpMRI images investigated in this study. Additionally, 2D CNNs are less computationally demanding and offer more flexibility in clinical settings. For these reasons, we opted to implement and evaluate the proposed framework in a 2D setting in the current study. Future studies should investigate whether a 3D implementation, with the rejection network providing a quality score for each slice in a 3D volume, could enhance segmentation accuracy.

The proposed method only generates ROI bounding boxes for interactive segmentation. Further development of the algorithm is needed to support other types of interactions, such as clicks, scribbles, and masks. The segmentation performance associated with these interaction types is required to be thoroughly assessed in future studies.

7. Conclusions

We have developed a robust framework that effectively integrates automatic segmentation with user interaction to enhance the performance and cost-effectiveness of image segmentation tasks. While existing networks such as MIMeSeg [42] and SAM [22] have incorporated human interaction to improve the accuracy of automatic segmentation, they are limited by the necessity of providing annotations for all input images, which diminishes their cost-effectiveness. Our primary innovation addresses this limitation by feeding an ROI, automatically extracted by a coarse segmentation network, into an interactive deep network to generate the final segmentation. However, because the automatic segmentation network has a limited ability to localize small and irregular-shaped lesions correctly, we proposed a rejection network to flag low-quality boundaries segmented by the coarse segmentation network for user interaction. We found substantial improvement by flagging about 20% of the images with the lowest quality score for manual annotation. With approximately 50% of the images manually annotated, the final segmentation accuracy was statistically indistinguishable from that achieved with full manual annotation. While this paper demonstrates our approach in prostate lesion segmentation from multimodality MRI, we believe it could be adapted to improve segmentation efficiency and maintain high accuracy standards in other medical image segmentation applications. Future evaluations involving image benchmarks of different organs are needed to assess the generalizability of the proposed framework.

Author Contributions: Conceptualization, W.K. and B.C.; methodology, W.K.; software, W.K.; validation, W.K.; formal analysis, W.K. and B.C.; investigation, W.K., C.R., H.M. and B.C.; resources, C.R., H.M. and B.C.; data curation, W.K., C.R. and H.M.; writing—original draft preparation, W.K.; writing—review and editing, W.K., C.R., H.M. and B.C.; visualization, W.K.; supervision, B.C.; project administration: B.C.; funding acquisition, B.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Research Grant Council of HKSAR, China (Project No. CityU 11205421).

Institutional Review Board Statement: The Prostate158 and PROSTATEx2 datasets used in our study are publicly available datasets that anonymize patient information. All contributions were approved by the institutional ethics committees as stated in [23,33–35].

Informed Consent Statement: The informed consents are not required because this is an investigation that utilizes the publicly available anonymized data.

Data Availability Statement: The data present in this study are publicly available in <https://zenodo.org/records/6481141>, <https://zenodo.org/records/6592345>, <https://www.aapm.org/GrandChallenge/PROSTATEx-2> (accessed on 1 August 2024), reference number [23,33–35]. Manual annotations of lesions from the Prostatex2 database are available from the corresponding author on reasonable request.

Acknowledgments: Bernard Chiu is grateful for the funding support from the Research Grant Council of HKSAR, China (Project No. CityU 11205421).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Siegel, R.L.; Miller, K.D.; Wagle, N.S.; Jemal, A. Cancer statistics, 2023. *CA Cancer J. Clin.* **2023**, *73*, 17–48. [CrossRef]
2. Olajide, A.O.; Eziyi, A.K.; Kolawole, O.A.; Sabageh, D.O.; Ajayi, I.A.; Olajide, F.O. A comparative study of the relevance of digital rectal examination, transrectal ultrasound, and prostate-specific antigen in the diagnostic evaluation of patients with advanced carcinoma of the prostate in a resource poor environment. *Sahel Med. J.* **2016**, *19*, 27.
3. Kasivisvanathan, V.; Rannikko, A.S.; Borghi, M.; Panebianco, V.; Mynderse, L.A.; Vaarala, M.H.; Briganti, A.; Budäus, L.; Hellawell, G.; Hindley, R.G.; et al. MRI-targeted or standard biopsy for prostate-cancer diagnosis. *N. Engl. J. Med.* **2018**, *378*, 1767–1777. [CrossRef] [PubMed]
4. Van der Leest, M.; Cornel, E.; Israel, B.; Hendriks, R.; Padhani, A.R.; Hoogenboom, M.; Zamecnik, P.; Bakker, D.; Setiasti, A.Y.; Veltman, J.; et al. Head-to-head comparison of transrectal ultrasound-guided prostate biopsy versus multiparametric prostate resonance imaging with subsequent magnetic resonance-guided biopsy in biopsy-naïve men with elevated prostate-specific antigen: A large prospective multicenter clinical study. *Eur. Urol.* **2019**, *75*, 570–578. [PubMed]
5. Weinreb, J.C.; Barentsz, J.O.; Choyke, P.L.; Cornud, F.; Haider, M.A.; Macura, K.J.; Margolis, D.; Schnall, M.D.; Shtern, F.; Tempany, C.M.; et al. PI-RADS Prostate Imaging – Reporting and Data System: 2015, Version 2. *Eur. Urol.* **2016**, *69*, 16–40. [CrossRef]
6. Steiger, P.; Thoeny, H.C. Prostate MRI based on PI-RADS version 2: How we review and report. *Cancer Imaging* **2016**, *16*, 9. [CrossRef] [PubMed]
7. Turkbey, B.; Rosenkrantz, A.B.; Haider, M.A.; Padhani, A.R.; Villeirs, G.; Macura, K.J.; Tempany, C.M.; Choyke, P.L.; Cornud, F.; Margolis, D.J.; et al. Prostate Imaging Reporting and Data System Version 2.1: 2019 Update of Prostate Imaging Reporting and Data System Version 2. *Eur. Urol.* **2019**, *76*, 340–351. [CrossRef]
8. Purysko, A.S.; Rosenkrantz, A.B.; Turkbey, I.B.; Macura, K.J. RadioGraphics update: PI-RADS version 2.1—A pictorial update. *Radiographics* **2020**, *40*, E33–E37. [CrossRef]
9. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
10. Isensee, F.; Jaeger, P.F.; Kohl, S.A.; Petersen, J.; Maier-Hein, K.H. nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* **2021**, *18*, 203–211. [CrossRef]
11. Wang, Z.; Liu, C.; Cheng, D.; Wang, L.; Yang, X.; Cheng, K.T. Automated detection of clinically significant prostate cancer in mp-MRI images based on an end-to-end deep neural network. *IEEE Trans. Med. Imaging* **2018**, *37*, 1127–1139. [CrossRef]
12. Arif, M.; Schoots, I.G.; Castillo Tovar, J.; Bangma, C.H.; Krestin, G.P.; Roobol, M.J.; Niessen, W.; Veenland, J.F. Clinically significant prostate cancer detection and segmentation in low-risk patients using a convolutional neural network on multi-parametric MRI. *Eur. Radiol.* **2020**, *30*, 6582–6592. [CrossRef] [PubMed]
13. Hambarde, P.; Talbar, S.; Mahajan, A.; Chavan, S.; Thakur, M.; Sable, N. Prostate lesion segmentation in MR images using radiomics based deeply supervised U-Net. *Biocybern. Biomed. Eng.* **2020**, *40*, 1421–1435. [CrossRef]
14. Qian, Y.; Zhang, Z.; Wang, B. ProCDet: A new method for prostate cancer detection based on mr images. *IEEE Access* **2021**, *9*, 143495–143505. [CrossRef]

15. Seetharaman, A.; Bhattacharya, I.; Chen, L.C.; Kunder, C.A.; Shao, W.; Soerensen, S.J.; Wang, J.B.; Teslovich, N.C.; Fan, R.E.; Ghanouni, P.; et al. Automated detection of aggressive and indolent prostate cancer on magnetic resonance imaging. *Med. Phys.* **2021**, *48*, 2960–2972. [[CrossRef](#)] [[PubMed](#)]
16. Zhang, G.; Shen, X.; Zhang, Y.D.; Luo, Y.; Luo, J.; Zhu, D.; Yang, H.; Wang, W.; Zhao, B.; Lu, J. Cross-modal prostate cancer segmentation via self-attention distillation. *IEEE J. Biomed. Health Inform.* **2021**, *26*, 5298–5309. [[CrossRef](#)] [[PubMed](#)]
17. Mehralivand, S.; Yang, D.; Harmon, S.A.; Xu, D.; Xu, Z.; Roth, H.; Masoudi, S.; Sanford, T.H.; Kesani, D.; Lay, N.S.; et al. A cascaded deep learning-based artificial intelligence algorithm for automated lesion detection and classification on biparametric prostate magnetic resonance imaging. *Acad. Radiol.* **2022**, *29*, 1159–1168. [[CrossRef](#)] [[PubMed](#)]
18. Pellicer-Valero, O.J.; Marenco Jimenez, J.L.; Gonzalez-Perez, V.; Casanova Ramon-Borja, J.L.; Martín García, I.; Barrios Benito, M.; Pelechano Gomez, P.; Rubio-Briones, J.; Rupérez, M.J.; Martín-Guerrero, J.D. Deep learning for fully automatic detection, segmentation, and Gleason grade estimation of prostate cancer in multiparametric magnetic resonance images. *Sci. Rep.* **2022**, *12*, 2975. [[CrossRef](#)] [[PubMed](#)]
19. Jiang, M.; Yuan, B.; Kou, W.; Yan, W.; Marshall, H.; Yang, Q.; Syer, T.; Punwani, S.; Emberton, M.; Barratt, D.C.; et al. Prostate cancer segmentation from MRI by a multistream fusion encoder. *Med. Phys.* **2023**, *50*, 5489–5504. [[CrossRef](#)]
20. Villers, A.; Lemaitre, L.; Haffner, J.; Puech, P. Current status of MRI for the diagnosis, staging and prognosis of prostate cancer: Implications for focal therapy and active surveillance. *Curr. Opin. Urol.* **2009**, *19*, 274–282. [[CrossRef](#)]
21. Zhao, F.; Xie, X. An overview of interactive medical image segmentation. *Ann. BMVA* **2013**, *2013*, 1–22.
22. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A.C.; Lo, W.Y.; et al. Segment anything. *arXiv* **2023**, arXiv:2304.02643.
23. Litjens, G.; Debats, O.; Barentsz, J.; Karssemeijer, N.; Huisman, H. Computer-aided detection of prostate cancer in MRI. *IEEE Trans. Med. Imaging* **2014**, *33*, 1083–1092. [[CrossRef](#)] [[PubMed](#)]
24. Zhou, T.; Li, L.; Bredell, G.; Li, J.; Unkelbach, J.; Konukoglu, E. Volumetric memory network for interactive medical image segmentation. *Med. Image Anal.* **2023**, *83*, 102599. [[CrossRef](#)] [[PubMed](#)]
25. Jiang, B.; Luo, R.; Mao, J.; Xiao, T.; Jiang, Y. Acquisition of localization confidence for accurate object detection. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 784–799.
26. Huang, Z.; Huang, L.; Gong, Y.; Huang, C.; Wang, X. Mask Scoring R-CNN. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 6409–6418.
27. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
28. Eidex, Z.A.; Wang, T.; Lei, Y.; Axente, M.; Akin-Akintayo, O.O.; Ojo, O.A.A.; Akintayo, A.A.; Roper, J.; Bradley, J.D.; Liu, T.; et al. MRI-based prostate and dominant lesion segmentation using cascaded scoring convolutional neural network. *Med. Phys.* **2022**, *49*, 5216–5224. [[CrossRef](#)] [[PubMed](#)]
29. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 19–23 June 2018; pp. 7132–7141.
30. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 19–23 June 2018; pp. 7794–7803.
31. Tan, M.; Le, Q. EfficientNet: Rethinking model scaling for convolutional neural networks. In Proceedings of the International Conference on Machine Learning, PMLR, Long Beach, CA, USA, 9–15 June 2019; pp. 6105–6114.
32. Ma, J.; He, Y.; Li, F.; Han, L.; You, C.; Wang, B. Segment anything in medical images. *Nat. Commun.* **2024**, *15*, 654. [[CrossRef](#)] [[PubMed](#)]
33. Adams, L.C.; Makowski, M.R.; Engel, G.; Rattunde, M.; Busch, F.; Asbach, P.; Niehues, S.M.; Vinayahalingam, S.; van Ginneken, B.; Litjens, G.; et al. Prostate158-An expert-annotated 3T MRI dataset and algorithm for prostate cancer detection. *Comput. Biol. Med.* **2022**, *148*, 105817. [[CrossRef](#)] [[PubMed](#)]
34. Clark, K.; Vendt, B.; Smith, K.; Freymann, J.; Kirby, J.; Koppel, P.; Moore, S.; Phillips, S.; Maffitt, D.; Pringle, M.; et al. The Cancer Imaging Archive (TCIA): Maintaining and operating a public information repository. *J. Digit. Imaging* **2013**, *26*, 1045–1057. [[CrossRef](#)]
35. Litjens, G.; Debats, O.; Barentsz, J.; Karssemeijer, N.; Huisman, H. ProstateX challenge data. *Cancer Imaging Arch.* **2017**, *10*, K9TCIA.
36. Yan, W.; Yang, Q.; Syer, T.; Min, Z.; Punwani, S.; Emberton, M.; Barratt, D.; Chiu, B.; Hu, Y. The impact of using voxel-level segmentation metrics on evaluating multifocal prostate cancer localisation. In Proceedings of the International Workshop on Applications of Medical AI, Singapore, 18 September 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 128–138.
37. Le Nobin, J.; Rosenkrantz, A.B.; Villers, A.; Orczyk, C.; Deng, F.M.; Melamed, J.; Mikheev, A.; Rusinek, H.; Taneja, S.S. Image guided focal therapy for magnetic resonance imaging visible prostate cancer: Defining a 3-dimensional treatment margin based on magnetic resonance imaging histology co-registration analysis. *J. Urol.* **2015**, *194*, 364–370. [[CrossRef](#)]
38. Gibson, E.; Bauman, G.S.; Romagnoli, C.; Cool, D.W.; Bastian-Jordan, M.; Kassam, Z.; Gaed, M.; Moussa, M.; Gómez, J.A.; Pautler, S.E.; et al. Toward prostate cancer contouring guidelines on magnetic resonance imaging: Dominant lesion gross and clinical target volume coverage via accurate histology fusion. *Int. J. Radiat. Oncol. Biol. Phys.* **2016**, *96*, 188–196. [[CrossRef](#)]
39. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. Pytorch: An imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 8026–8037.

40. Loshchilov, I.; Hutter, F. Decoupled weight decay regularization. *arXiv* **2017**, arXiv:1711.05101.
41. Hatamizadeh, A.; Tang, Y.; Nath, V.; Yang, D.; Myronenko, A.; Landman, B.; Roth, H.R.; Xu, D. UNETR: Transformers for 3D medical image segmentation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2022; pp. 574–584.
42. Luo, X.; Wang, G.; Song, T.; Zhang, J.; Aertsen, M.; Deprest, J.; Ourselin, S.; Vercauteren, T.; Zhang, S. MIDeepSeg: Minimally interactive segmentation of unseen objects from medical images using deep learning. *Med. Image Anal.* **2021**, *72*, 102102. [[CrossRef](#)] [[PubMed](#)]
43. David, H.A.; Gunnink, J.L. The paired t test under artificial pairing. *Am. Stat.* **1997**, *51*, 9–12. [[CrossRef](#)]
44. Rosenkrantz, A.B.; Ayoola, A.; Hoffman, D.; Khasgiwala, A.; Prabhu, V.; Smereka, P.; Somberg, M.; Taneja, S.S. The learning curve in prostate MRI interpretation: Self-directed learning versus continual reader feedback. *Am. J. Roentgenol.* **2017**, *208*, W92–W100. [[CrossRef](#)]
45. Lee, G.H.; Chatterjee, A.; Karademir, I.; Engelmann, R.; Yousuf, A.; Giurcanu, M.; Harmath, C.B.; Karczmar, G.S.; Oto, A. Comparing radiologist performance in diagnosing clinically significant prostate cancer with multiparametric versus hybrid multidimensional MRI. *Radiology* **2022**, *305*, 399–407. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.