Genitourinary Imaging • Original Research

Three-Dimensional Convolutional Neural Network for Prostate MRI Segmentation and Comparison of Prostate Volume Measurements by Use of Artificial Neural Network and Ellipsoid Formula

Dong Kyu Lee[1]
Deuk Jae Sung[1]
Chang-Su Kim[2]
Yuk Heo[2]
Jeong Yoon Lee[1]
Beom Jin Park[1]
Min Ju Kim[1]

OBJECTIVE. The purposes of this study were to assess the performance of a 3D convolutional neural network (CNN) for automatic segmentation of prostates on MR images and to compare the volume estimates from the 3D CNN with those of the ellipsoid formula.

MATERIALS AND METHODS. The study included 330 MR image sets that were divided into 260 training sets and 70 test sets for automated segmentation of the entire prostate. Among these, 162 training sets and 50 test sets were used for transition zone segmentation. Assisted by manual segmentation by two radiologists, the following values were obtained: estimates of ground-truth volume (V_{GT}), software-derived volume (V_{SW}), mean of V_{GT} and V_{SW} (V_{AV}), and automatically generated volume from the 3D CNN (V_{NET}). These values were compared with the volume calculated with the ellipsoid formula (V_{EL}).

RESULTS. The Dice similarity coefficient for the entire prostate was 87.12% and for the transition zone was 76.48%. There was no significant difference between V_{NET} and V_{AV} ($p = 0.689$) in the test sets of the entire prostate, whereas a significant difference was found between V_{EL} and V_{AV} ($p < 0.001$). No significant difference was found among the volume estimates in the test sets of the transition zone. Overall intraclass correlation coefficients between the volume estimates were excellent (0.887–0.995). In the test sets of entire prostate, the mean error between V_{GT} and V_{NET} (2.5) was smaller than that between V_{GT} and V_{EL} (3.3).

CONCLUSION. The fully automated network studied provides reliable volume estimates of the entire prostate compared with those obtained with the ellipsoid formula. Fast and accurate volume measurement by use of the 3D CNN may help clinicians evaluate prostate disease.

The volume of the prostate has clinical significance for treatment planning and posttreatment follow-up of prostate disorders. Accurate measurement of prostate volume is required to calculate prostate-specific antigen density (PSAD), which is the ratio of serum PSA level to prostate volume. PSAD is helpful in stratifying risk of prostate cancer and selecting candidates for active surveillance [1–5]. Among the methods for prostate volume estimation, the ellipsoid formula based on transrectal ultrasound (TRUS) measurements is the most common approach [4–6]. Although TRUS is an easily accessible bedside imaging modality, it is highly operator dependent and is vulnerable to artifacts. Given that MRI allows more objective volume measurement and yields higher-resolution images than TRUS does, MRI is preferable to TRUS for determination of prostate volume [5, 7–9]. The configurations of individual prostates vary, and the ellipsoid formula is based on the assumption that the prostate is completely ellipsoid. As a result, there is a significant difference between actual prostate volume and the volume calculated with the ellipsoid formula on MR images.

Manual planimetry can be used to avoid potential inaccuracy caused by the variable contour of the prostate [8, 9]. Image processing software can be used to derive an estimated prostate volume from the integration of manually segmented areas for all slices. Despite the higher accuracy and reliability of this method, it is time-consuming and thus difficult to routinely implement in daily clinical practice. Therefore, fully automated 3D segmentation for prostate volume estimation has attracted great interest. In particular, the live challenge workshop Prostate MR Image Segmentation 2012 (Promise12) was held at

the 2012 International Conference on Medical Image Computing and Computer Assisted Intervention to compare the performance of automatic segmentation algorithms for prostate MRI [10]. A total of 100 prostate MR images from different centers was assembled into a set of 50 datasets for training, 30 datasets for testing, and 20 datasets for the live challenge with the corresponding ground-truth regions of the prostate.

Several groups have evaluated their proposed deep learning models with these open-source datasets [11–13]. Other prior prostate segmentation networks, called V-net and baseline convolutional neural network (CNN), provided motivation for this study [14, 15]. V-net is a fully convolutional neural network for prostate segmentation and was trained with the Promise12 challenge dataset. V-net performance was also evaluated with 30 Promise12 test datasets. Like V-net, baseline CNN was trained with the Promise12 training dataset, and its performance was compared with several objective functions on the Promise12 dataset. Inspired by the Promise12 challenge, we introduce our new prostate MRI dataset, namely KUProstate, KU indicating Korea University. We also propose a 3D CNN for two-class segmentation. After describing our network and method, we assess the performance of our network and compare the results with conventional volume measurements obtained with the ellipsoid formula.

Materials and Methods
Patients and Study Design

This single-institution, retrospective study was approved by the institutional review board, and the requirement for informed consent was waived. Our newly proposed dataset (KUProstate) is composed of 330 prostate MR image sets from patients treated at our university hospital between January 2008 and January 2017. The 330 prostate MR image sets were divided into 260 training sets and 70 test sets for our network. Labeled data were generated for the entire prostate. For the transition zone, we used 162 MR image sets for training and 50 sets for testing. A total of 118 cases were excluded because of difficulty in clearly delineating the transition zone (e.g., anatomic abnormalities due to prostate cancer, hemorrhage, or small prostate volume).

Manual segmentation of the entire prostate and transition zone boundaries was performed by two radiologists, each with 4 years of experience in prostate MRI. ROIs were manually drawn on each slice for network training, and the integrated prostate volume was simultaneously selected as ground truth for our volume data (V_{GT}). In ad-

dition, measurement of the prostate volume with software (Aquarius 3D Workstation, TeraRecon) was repeated in all cases after 1 month (V_{SW}). After training the network, we obtained volume estimates that were automatically generated from the network operating on the test datasets (V_{NET}).

To compare the performance of our network with the outcome of conventional volume determination, the volume estimates of all prostates were calculated by the ellipsoid formula (V_{EL} = transverse diameter × anteroposterior diameter × height × 0.52) [16]. The maximum anteroposterior and transverse dimensions were manually measured on axial T2-weighted images, and the height of the prostate was measured on sagittal T2-weighted images.

MRI Technique

Prostate MR images were acquired with a Tim Trio 3-T Magnetom system (Siemens Healthineers) with a pelvic phased-array coil without an endorectal coil. Among the sequences included in the standard protocol, axial and sagittal T2-weighted turbo spin-echo images were used for prostate segmentations and volume determinations. The parameters of the T2-weighted images were as follows: FOV, 200 × 200 to 270 × 270 mm; matrix, 384–448 × 269–384; TR/TE, 6020–7430/91–135; flip angle, 120–140°; slice thickness, 3 mm; number of signals acquired, 3–4; echo-train length, 15–49; voxel size, 0.4–0.7 × 0.4–0.7 × 3.0 mm.

Proposed Network Architecture and Implementation

The aim of our proposed network is to infer the 3D prostate voxel image that matches the annotation in the entire prostate and the transition zone. To complete this task, before using the network, we first resized the 3D MR input images into 3D voxel images. The voxel images had 192 × 192 × 48 voxels each.

Inspired by V-net and baseline CNN, our proposed network consisted of an encoding-decoding method with convolution and deconvolution. As shown in Figure 1, a single encoding block comprised four convolution layers. In a block, there were three convolution layers and a downsampling layer. To maintain symmetry, a single decoding block comprised an upsampling layer and three convolution layers. In the bridge, the data went through three convolution layers to obtain various features from a broad region. Each convolution layer was a 3 × 3 × 3 filter layer. In the downsampling layer, we used a 2 × 2 × 2 filter convolution layer with stride 2 to obtain global and representative features of the input image. By using convolution rather than a pooling layer for downsampling, our proposed network effectively eliminated use-

less features. In the upsampling layer, we used deconvolution layers [17].

The coefficients of the filter of the deconvolution layer were determined through training. The deconvolution layer with hierarchic structure could register detailed information at various scales. We used a parametric rectified linear unit as an activation function. Batch normalization was applied to deal with the problem of vanishing and exploding gradients as the layers become deeper [18, 19]. The activation function and batch normalization were applied in this order to each layer output.

Simplified residual additions (gray arrows, Fig. 1) were one of the emphases of this study. Although our network shared many common advantages with the baseline CNN for featuring a similar network structure, the simplification in the fine-to-coarse residual additions improved performance on segmentation. We exploited the residual feature forwarding strategy for fine data scale in decoding and for coarse data scale at image classification level, as V-net does. Thus, we could use a variety of information from the data [15, 20]. Especially in the transition zone segmentation, our symmetric residual additions constituted a necessary strategy because the segmentation region was much smaller than the existing region from which we needed to exploit fine information from the data.

We augmented the training dataset by applying flipping with a probability of 0.5. As typically performed in CNN, we randomly cropped 96 × 96 × 16 subvolume patches with stride 1 from 192 × 192 × 48 to enlarge the data [21]. For the objective function, we used the cross-entropy loss defined as follows:

$$L_c = -\sum_{i=1}^{N} q_i \log(p_i) - \sum_{i=1}^{N}(1 - q_i)\log(1 - p_i),$$

where N represents the total number of voxels in the output image and p_i and q_i denote the probability that the i-th voxel of the ground truth corresponds to the respective prostate region of the network segmentation result.

Given that we used a new dataset, for unity, we used only the cross-entropy loss function for training loss on the other methods already proposed for comparison. Our batch size was fixed at 12, and we used the Adam optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$) to optimize the cost [22]. The initial learning rate was set to 0.005 with one-fourth decay every 20,000 batches in the entire region segmentation. Concerning the transition region segmentation, given that it relied on approximately one-half of the data compared with the entire prostate, the batches for decay were also halved. We used the TensorFlow open source machine learning platform (Google Brain Team, Apache License 2.0) to implement the entire experiment [23]. The process for the entire pros-
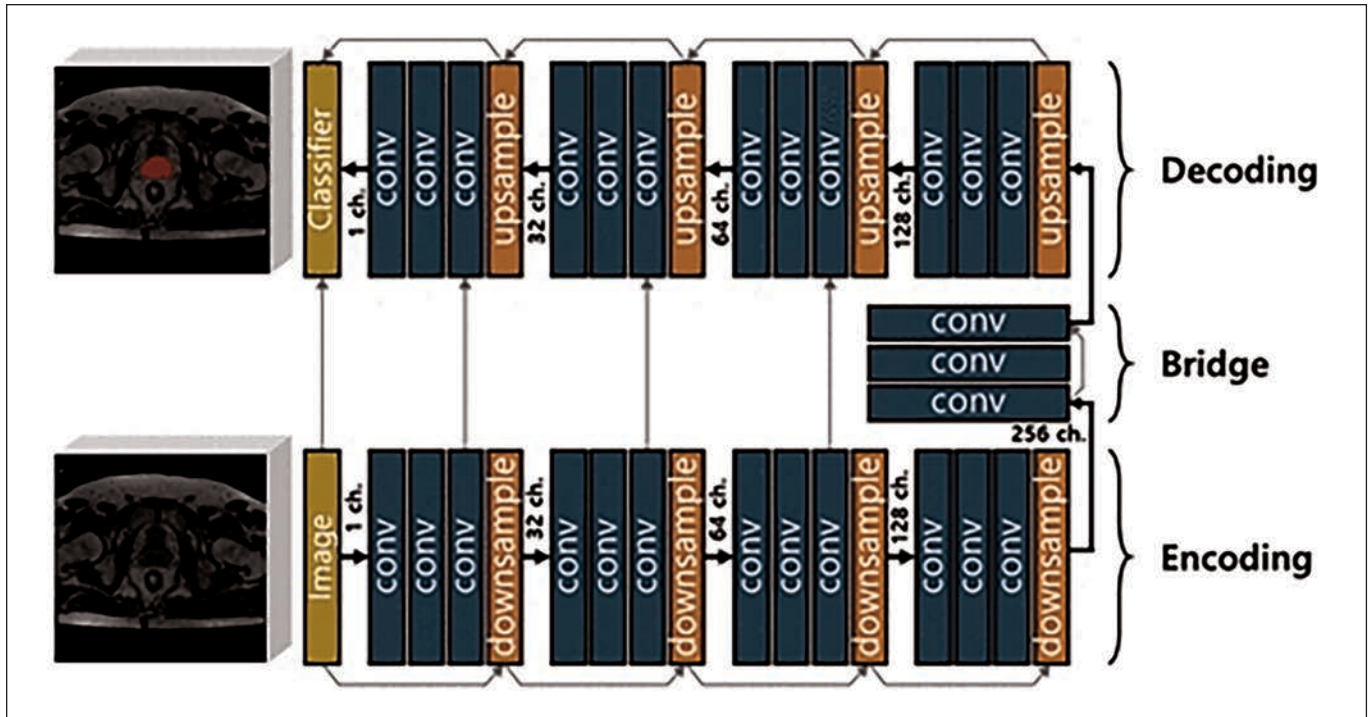
Fig. 1—Schematic shows network architecture. Each convolution layer is $3 \times 3 \times 3$ filter. Gray arrows denote residual additions. ch. = channel or channels, conv = convolution.

tate took approximately 5 hours and for the transition region, approximately 2.5 hours with a 1080-Ti graphics processing unit (GeForce GTX, Nvidia).

Volume Data Acquisition

As V_{EL} was calculated with the ellipsoid formula, we obtained V_{GT} and V_{NET} in an integral manner. Typically, a single ground-truth volume in a prostate MR image is acquired by integrating the approximated lines that are annotated manually for every slice. From the approximated analog lines, we created a digitalized 3D shape of the prostate. We then integrated the voxels considering *xyz*-spacing from the digital prostate 3D image to get the annotated volume. During the network process, we calculated the parameters to resize the original MR images to $192 \times 192 \times 48$ shaped images to use them as the inputs of the network. The calculated segmentation from the network then retrieved the original size through those resizing parameters. Finally, we obtained the volume of the segmentation by using the same integral approach as for acquiring the ground-truth volume.

Statistical Analysis

For evaluation of the segmentation performance of the proposed algorithm, we used the Dice similarity coefficient (S_D), which is commonly applied for prostate segmentation and is a score defined according to the following equation with the notations used in the previous equation:

$$S_D = \frac{2\sum_{i=1}^{N} p_i^2 q_i}{\sum_{i=1}^{N} p_i^2 + \sum_{i=1}^{N} q_i^2}.$$

The Dice similarity coefficient provided the degree of spatial agreement between manual segmentation by the radiologists and automatic segmentation by the proposed network. It ranged between 0.0 (no overlap at all between the segmentations) and 1.0 (perfect match); greater values indicated better spatial agreement between the segmentations.

We used SPSS software (SPSS Statistics 22.0, IBM) for all statistical analyses. To compare the differences and assess the correlations between V_{GT}, V_{SW}, V_{EL}, and V_{NET} using data from both the entire and transition zones, we used the Wilcoxon signed rank test and intraclass correlation coefficients (ICCs). Comparison of the measurement errors among the methods was performed with Bland-Altman plots and box plots in each dataset. To compare the accuracy of volume estimates from our network (V_{NET}) with the accuracy of conventional volume measurement using the ellipsoid formula (V_{EL}), the mean or median of several measurement errors, including error (error$_1$ = $V_{GT} - V_{EL}$, error$_2$ = $V_{GT} - V_{NET}$), squared error (squared error$_1$ = $[V_{GT} - V_{EL}]^2$, squared error$_2$ = $[V_{GT} - V_{NET}]^2$), absolute percentage error (absolute percentage error$_1$ = $| [V_{GT} - V_{EL}] / V_{GT} | \times 100$, absolute percentage error$_2$ = $| [V_{GT} - V_{NET}] / V_{GT} | \times 100$), and absolute error (absolute error$_1$ = |

$V_{GT} - V_{EL}$ |, absolute error$_2$ = | $V_{GT} - V_{NET}$ |), were compared with each other in both datasets.

Results

Using the Dice similarity coefficient, we quantitatively compared the segmentation utility of our proposed algorithm with the other two methods considered regarding two different prostate regions. Our proposed algorithm had the largest Dice similarity coefficient values in both the entire prostate and the transition zone (Table 1). The Dice

TABLE 1: Quantitative Comparison Between Proposed Algorithm and the V-Net and Baseline Convolutional Neural Network (CNN) Methods

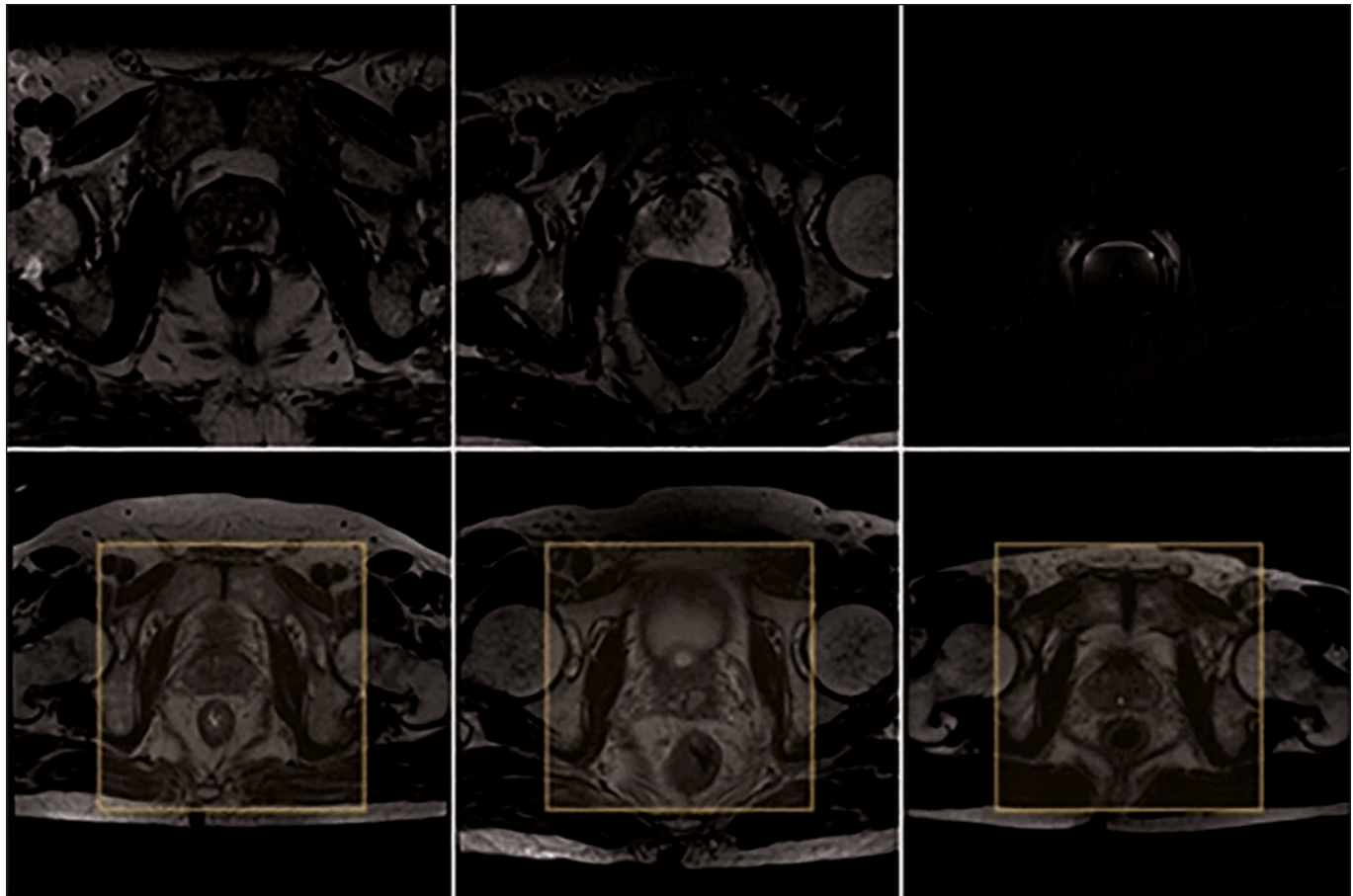| Method	Dice Coefficient (%)
Entire prostate	
V-net	63.02
Baseline CNN	78.13
Proposed	87.12
Transition zone of prostate	
V-net	64.15
Baseline CNN	70.57
Proposed	76.48

Fig. 2—MR images show three slices of data. Yellow box indicates physical area corresponding to Promise12 network (*top*) for comparison with physical area of KUProstate network (*bottom*).

similarity coefficient achieved by the proposed algorithm for the entire prostate was 87.12%. The result in the transition zone was lower (76.48%). Unlike the other two methods, our network used a larger voxel image and subvolume training (Fig. 2). This difference caused our algorithm to perform better. Figure 3 shows two slices of prostate segmentation in the entire prostate and the transition zone and the resulting performance of our algorithm. The figure also allows size comparison between entire prostates and transition zones in our dataset.

Baseline characteristics of the training and test datasets in the entire prostate and the transition zone included mean, SD, and median values of all volume estimates (V_{GT}, V_{SW}, V_{EL}, and V_{NET}) and mean of V_{GT} and V_{SW} (V_{AV}). The ellipsoid error rate was defined as the mean absolute percentage error between V_{GT} and V_{EL}, and the network error rate was the mean absolute percentage error between V_{GT} and V_{NET}. In the training dataset of the entire prostate, there was

little difference between the means of V_{GT}, V_{SW}, and V_{EL} and V_{AV}. Likewise, the SDs of all volume estimates exhibited little difference (Table 2). In the test dataset of the entire prostate, the mean and median of V_{GT} were minimally larger than those of the other volume estimates. These differences seemed too small to have clinical significance. The differences between the values of all volume estimates were also negligible in the training and test datasets of the transition zone.

The Wilcoxon signed rank test revealed no statistically significant difference between the volume estimates in the training dataset of the entire prostate (Table 3). In contrast, there were significant differences between V_{GT}, V_{EL}, and V_{NET} in the test dataset of the entire prostate. There was no statistically significant difference between V_{NET} and V_{AV} ($p = 0.689$), but there was a significant difference between V_{EL} and V_{AV} ($p < 0.001$). In the training dataset of the transition zone, there were significant differences between V_{GT} and the other volume estimates ($p <$

0.001), corresponding to systematic error. No significant difference was found between the volume estimates in the test dataset of the transition zone. Overall ICCs between the volume estimates were excellent in both the entire prostate and the transition zone, with a slight decrease in ICC between V_{NET} (0.890) and V_{GT} (or V_{AV}) (0.877) in the test dataset of the transition zone.

In the test dataset of the entire prostate, the mean error between V_{GT} and V_{NET} (2.52) was smaller than that between V_{GT} and V_{EL} (3.34). The mean squared error between V_{GT} and V_{NET} (37.01) was larger than that between V_{GT} and V_{EL} (28.29), which was due to a larger SD of the error between V_{GT} and V_{NET}. The median absolute error between V_{GT} and V_{NET} (2.30) was also smaller than that between V_{GT} and V_{EL} (3.04). Up to the 80 percentile, the absolute errors between V_{GT} and V_{NET} were lower than those between V_{GT} and V_{EL}. The mean absolute error between V_{GT} and V_{NET} (3.82) was also smaller than that between V_{GT} and V_{EL} (3.95) in
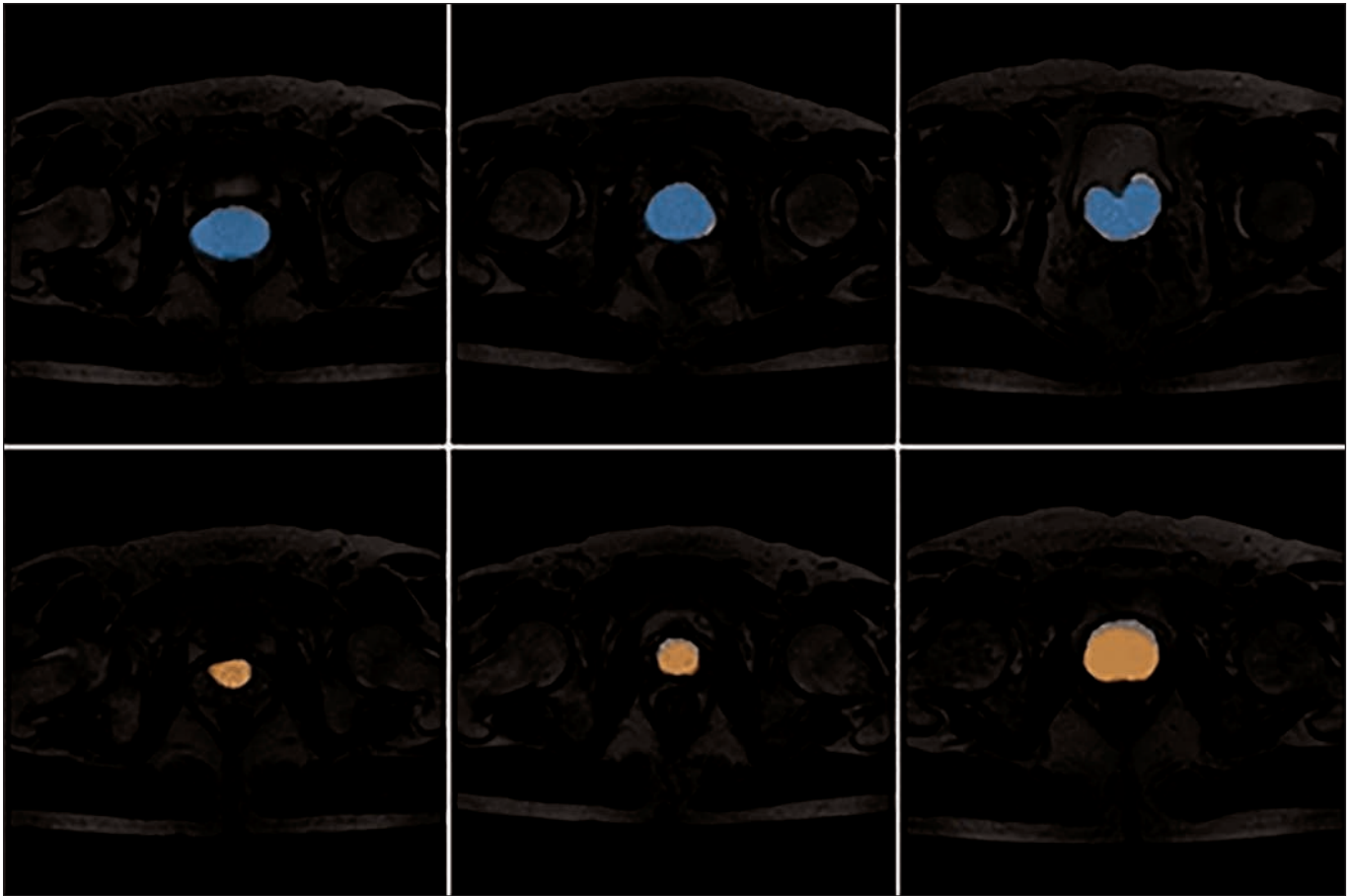
Fig. 3—Qualitative analysis of proposed algorithm in three test sets (*left to right*, sets 261, 279, and 288). MR images show entire prostate (*top*) and transition zone (*bottom*). Blue and orange denote ground truth in respective regions; gray surrounding ground truth denotes performance of proposed network.

the test sets of the entire prostate. By contrast, the mean absolute error between V_{GT} and V_{NET} (3.08) was slightly larger than that between V_{GT} and V_{EL} (1.78) in the test dataset of the transition zone. Comparison of measurement errors among the methods in the test datasets of both zones is summarized in Table 4.

Bland-Altman plots and box plots in all datasets are shown in Figures 4 and 5. Figures 4A and 4B show the poor fit between V_{GT} and V_{NET} with a large range between the upper and lower limits of agreement in the test dataset of the entire prostate, though the mean difference between V_{GT} and V_{NET} was closer to zero (2.5) than the mean difference between V_{GT} and V_{EL} (3.3). In the test dataset from the transition zone, the 95% CI of the limits of agreement between V_{GT} and V_{NET} was also wider than that between V_{GT} and V_{EL} (Figs. 4C and 4D). Figure 5 shows that the training datasets of V_{GT} for both the entire prostate and the transition zone did not include enough data with a large volume.

Discussion

Over the last several years, there have been broad applications of machine learning and artificial intelligence algorithms to imaging of the prostate. In addition to many automated segmentation models with prostate MRI, CNN-based prostate segmentation with PET/CT has been introduced [24, 25]. The range of clinical applications of neural networks to prostate MRI includes gland segmentation, tumor detection and localization, assessment of lesion aggressiveness, local staging and pretreatment assessment, and prediction of biochemical recurrence [26]. Most of these applications are based on T2-weighted MR images. In a specific setting such as prostate cancer classification, other MRI sequences, including apparent diffusion coefficient mapping, DWI, and perfusion imaging (K^{trans}), have been used to validate the effects of different sequences on performance of a proposed method [27]. Unfortunately, several challenges for accurate prostate segmenta-

tion continue to exist. These problems are mainly due to great variability in the size and shape of prostates and difficulty in separating the prostate from surrounding anatomic structures [12, 13, 28, 29]. Therefore, many previous studies have included manual segmentation by experienced radiologists as ground truth segmentation of prostates [25, 28–31].

In KUProstate, the entire prostate was annotated in the same manner as the already existing Promise12 dataset [10]. Our dataset, however, had differences from Promise12. First, the KUProstate prostate MR images contained less noise and had larger spatial resolution than those of Promise12, which means that our dataset contained more information for deep learning. In addition, Promise12 has currently annotated only 50 datasets, whereas KUProstate has annotated approximately 330 cases. Among them, 212 sets also contain the transition zone information. Table 5 compares the differences between Promise12 and KUProstate.

TABLE 2: Baseline Characteristics of the Training and Test Datasets in Entire Prostate and the Transition Zone

Variable	Entire Prostate			Transition Zone		
	Mean	SD	Median	Mean	SD	Median
Training dataset						
V_{GT}	35.83	15.56	32.24 (14.16–104.21)	17.64	12.99	12.69 (3.99–67.15)
V_{SW}	35.70	15.60	31.50 (14.60–101.00)	15.49	11.84	11.80 (2.92–66.70)
V_{EL}	35.88	15.48	32.98 (14.31–103.80)	14.74	10.14	12.27 (3.83–65.30)
V_{NET}						
V_{AV}	35.77	15.50	31.65 (14.80–102.61)	16.12	12.06	12.30 (4.00–66.93)
Test dataset						
V_{GT}	33.17	14.38	29.20 (11.80–81.13)	14.42	9.45	11.15 (2.63–42.79)
V_{SW}	30.44	13.13	27.15 (11.80–75.10)	14.11	9.02	11.10 (2.64–43.70)
V_{EL}	29.83	12.95	27.67 (10.33–78.86)	14.07	9.04	10.39 (3.11–42.70)
V_{NET}	30.64	12.72	27.72 (6.90–73.29)	15.18	11.47	10.90 (3.35–50.67)
V_{AV}	31.80	13.65	28.60 (11.80–78.11)	14.26	9.19	10.95 (2.64–42.68)
Ellipsoid error rate	11.92	8.45	9.63 (0.21–36.12)	13.83	9.55	12.41 (0.29–35.48)
Network error rate	11.78	13.30	7.37 (0.09–71.27)	25.17	36.09	15.22 (0.29–241.42)

Note—Values in parentheses are ranges. Prostate volume is measured in cubic centimeters. V_{GT} = ground-truth volume, V_{SW} = software-derived volume, V_{EL} = volume calculated with ellipsoid formula, V_{NET} = automatically generated volume from network, V_{AV} = mean of V_{GT} and V_{SW}.

Unlike studies conducted with conventional datasets, the current study provided annotations of the transition zones of prostates in 212 cases. Whereas previous studies, except for a 2019 report including central gland segmentation [31], have focused on the possibility of segmentation on prostate MR images, they have addressed only segmentation of the entire prostate. Promise12, which is a dataset of the entire prostate, has been the dataset most commonly used for prostate segmentation. To our knowledge, there has been neither a shared dataset nor any research in which algorithms have been studied. Prostate segmentation performance is currently focused on general segmentation performance in the conventional computer vision field. As CNN-based segmentation has improved, this method has superseded manual segmentation within an affordable error range. Nevertheless, except for Promise12, datasets have not expanded to encompass detailed segmentation of the prostate. In this study, we provided a new dataset for the transition zone of the prostate to solve this problem and to set a new goal in the study of stagnant prostate segmentation.

Although our results showed excellent correlation between V_{GT} and both V_{EL} and V_{NET}, the mean error between V_{GT} and V_{NET} was smaller than that between V_{GT} and V_{EL}. In the upper percentiles, however, the absolute errors between V_{GT} and V_{NET} were larger than those between V_{GT} and V_{EL}. Slightly wider limits of agreement were identified between V_{GT} and V_{NET} than between V_{GT} and V_{EL}. This suggests the absence of sufficient data with large volumes in the training dataset of the entire prostate. This data imbalance could be overcome by oversampling the large volume data.

Regarding the transition zone, segmentation was limited, and there was a degree of systematic error that could not be avoided. Prediction in this region was challenging because it required segmentation of a much smaller area than did the experiment in the entire prostate. In addition, a wider 95% CI of the limits of agreement was seen between V_{GT} and V_{NET} than between V_{GT} and V_{EL}. As in the training dataset of the entire prostate, the training dataset of V_{GT} for the transition zone did not have enough data with large volumes.

TABLE 3: Results of Wilcoxon Signed Rank Test and Intermethod Correlation

Comparison of Methods	Wilcoxon Signed Rank Test (p)		Intraclass Correlation Coefficient[a]	
	Entire Prostate	Transition Zone	Entire Prostate	Transition Zone
Training dataset				
V_{GT} vs V_{SW}	0.196	< 0.001	0.979	0.965
V_{GT} vs V_{EL}	0.883	< 0.001	0.926	0.925
V_{GT} vs V_{AV}	0.196	< 0.001	0.995	0.991
V_{EL} vs V_{AV}	0.657	< 0.001	0.933	0.943
Test dataset				
V_{GT} vs V_{SW}	< 0.001	0.357	0.948	0.979
V_{GT} vs V_{EL}	< 0.001	0.273	0.927	0.971
V_{GT} vs V_{NET}	0.001	0.431	0.901	0.890
V_{GT} vs V_{AV}	< 0.001	0.357	0.987	0.995
V_{EL} vs V_{AV}	< 0.001	0.286	0.964	0.986
V_{NET} vs V_{AV}	0.689	0.265	0.907	0.877

Note—V_{GT} = ground-truth volume, V_{SW} = software-derived volume, V_{EL} = volume calculated with ellipsoid formula, V_{AV} = mean of V_{GT} and V_{SW}, V_{NET} = automatically generated volume from the network.
[a]All $p < 0.001$.

TABLE 4: Comparison of Measurement Errors Among Methods in the Test Datasets of the Entire Prostate and Transition Zone

Measurement Error	Mean	SD	Median	Q1	Q3	P80	P90
Entire prostate							
$Error_1$	3.34	4.17	2.72	0.82	5.18	5.96	7.91
$Squared\ error_1$	28.29	60.32	9.24	1.91	26.82	35.54	62.59
$Absolute\ percentage\ error_1$	11.92	8.45	9.63	5.95	17.89	18.87	24.72
$Absolute\ error_1$	3.95	3.59	3.04	1.38	5.18	5.96	7.91
$Error_2$	2.52	5.58	0.96	−0.47	4.21	5.90	9.26
$Squared\ error_2$	37.01	92.82	5.31	0.83	24.37	35.01	85.72
$Absolute\ percentage\ error_2$	11.78	13.30	7.37	3.25	16.86	21.59	24.12
$Absolute\ error_2$	3.82	4.77	2.30	0.91	4.93	5.91	9.26
Transition zone							
$Error_1$	0.35	2.23	0.35	−1.08	2.17	2.32	5.43
$Squared\ error_1$	5.00	6.59	1.78	0.43	7.40	8.49	29.44
$Absolute\ percentage\ error_1$	13.83	9.55	12.41	6.75	19.42	20.22	35.48
$Absolute\ error_1$	1.78	1.37	1.33	0.66	2.72	2.91	5.43
$Error_2$	−0.77	4.92	−0.16	−2.28	1.38	1.92	14.29
$Squared\ error_2$	24.35	68.10	2.83	0.78	12.70	20.97	394.76
$Absolute\ percentage\ error_2$	25.17	36.09	15.22	7.96	34.94	39.17	241.42
$Absolute\ error_2$	3.08	3.89	1.68	0.89	3.56	4.58	19.87

Note—Q1 = first quartile, Q3 = third quartile, P80 = 80th percentile, P90 = 90th percentile.

Measurement of prostate volume can influence many clinical scenarios, especially in the management of prostate cancer. PSA measurement alone is not a specific surrogate marker in prostate cancer, especially at intermediate serum PSA levels between 4.0 and 10.0 ng/mL [32, 33]. Instead of serum PSA concentration, PSAD is a good differentiator between prostate cancer and benign prostatic disease. Furthermore, PSAD can be a useful monitoring tool in active surveillance for low-risk prostate cancer and can be a predictor of the risk of disease progression during follow-up [33, 34]. Accurate measurement of prostate volume is necessary to calculate PSAD. The ellipsoid formula, which has been traditionally used for prostate volume measurement, frequently results in er-
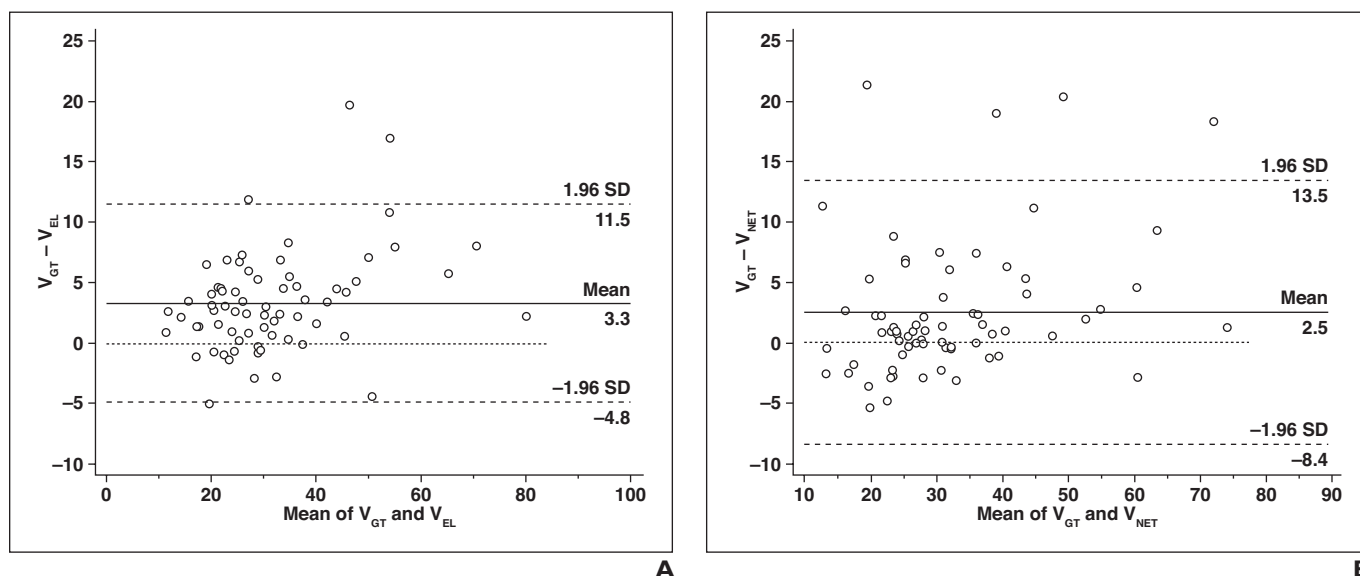

Fig. 4—Bland-Altman plots of test datasets. Prostate volume is measured in cubic centimeters. V_{GT} = volume data integrated from ground truth, V_{EL} = volume data integrated from ellipsoid formula, V_{NET} = volume data integrated from convolutional neural network.
A, Entire prostate results for mean V_{GT} and V_{EL}.
B, Entire prostate results for V_{GT} and V_{NET}.
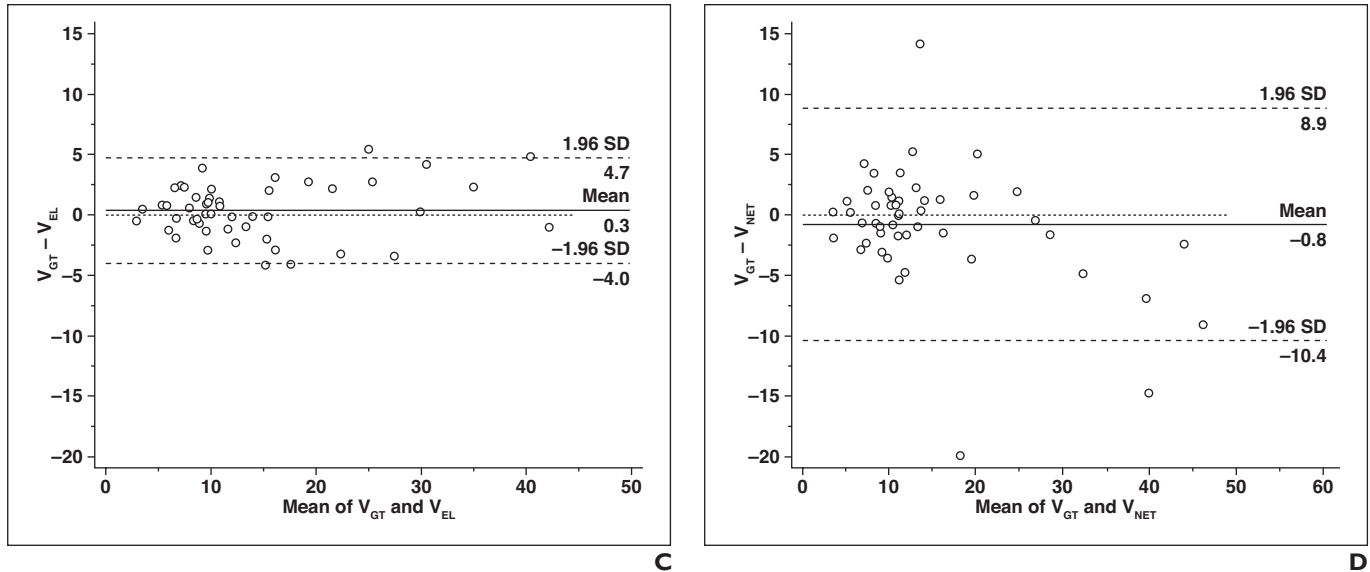
(Fig. 4 continues on next page)

Fig. 4 (continued)—Bland-Altman plots of test datasets. Prostate volume is measured in cubic centimeters. V_{GT} = volume data integrated from ground truth, V_{EL} = volume data integrated from ellipsoid formula, V_{NET} = volume data integrated from convolutional neural network.
C, Transition zone results for V_{GT} and V_{EL}.
D, Transition zone results for V_{GT} and V_{NET}.
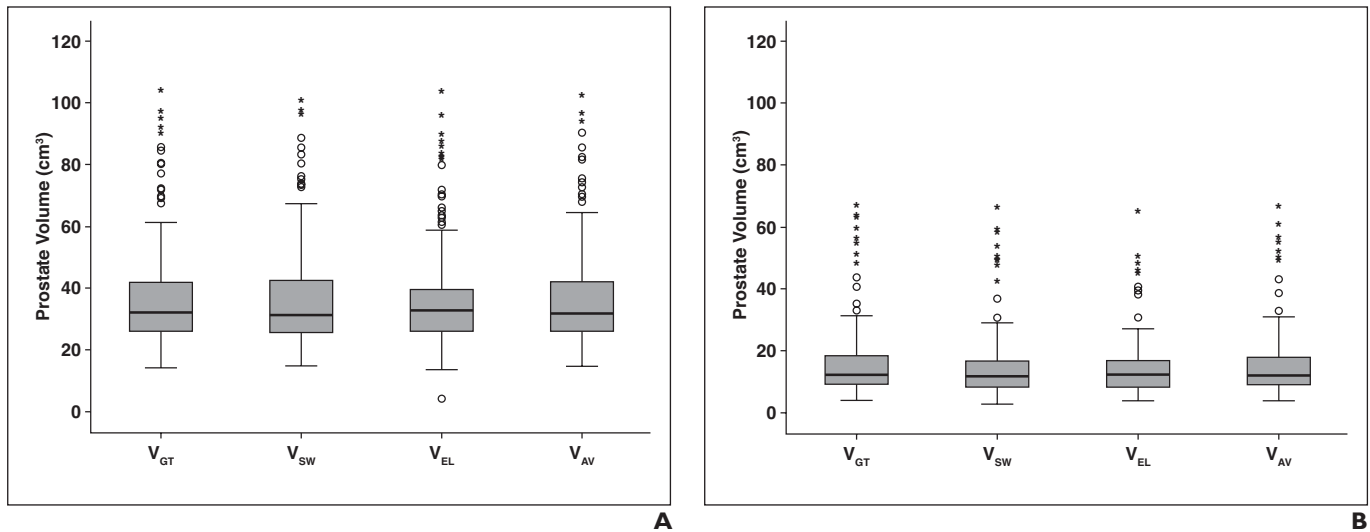

Fig. 5—Box plots of training datasets. Prostate volume is measured in cubic centimeters. Upper and lower limits of whiskers are values furthest from median but within distance to corresponding fence. Here, the lower and upper fences are defined as upper fence, 75th quartile + 1.5 interquartile range (IQR); lower fence, 25th quartile −1.5 IQR. V_{GT} = volume data integrated from ground truth, V_{SW} = volume data integrated from software, V_{EL} = volume data integrated from ellipsoid formula, V_{AV} = mean of V_{GT} and V_{SW}.
A, Entire prostate.
B, Transition zone.

rors if median lobe hypertrophy or asymmetry of the prostate is present [7, 33, 35].

Because the ellipsoid formula is based on the assumption that the prostate is completely ellipsoid, many actual prostates with hypertrophy or irregular contours do not satisfy this assumption. Moreover, even a small error in one diameter can affect the volume estimate with amplified error because the prostate volume is calculated by multiplying 3D parameters in the ellipsoid

formula [36]. More precise and reliable estimation of prostate volume can be achieved with manual planimetry [37]. However, manual segmentation using planimetry is extremely time-consuming and thus rarely performed in routine clinical practice. To overcome the limitations of the aforementioned conventional volume measurement methods, several semiautomated or fully automated tools for prostate volume determination with MRI have been developed,

but there are few data on the level of use of these techniques [4, 38, 39].

Attempts have been made to segment the prostate automatically at MRI by use of machine learning technology. Although there have been several prostate MRI datasets and neural networks associated with automatic segmentation, most of the investigators have compared their results with those obtained with other methods through Dice similarity coefficients only [40–42]. To our knowledge,

TABLE 5: Differences Between Prostate MR Image Segmentation 2012 (Promise12) and Korea University (KUProstate) Datasets

Characteristic	Promise12	KUProstate
No. of cases	100[a]	330[b]
Patients	Patients with both benign disease (e.g., benign prostatic hyperplasia) and prostate cancer	Patients with both benign disease (e.g., benign prostatic hyperplasia) and prostate cancer
DICOM file	Not available	Available
Field strength (T)	1.5 or 3.0	3.0
Endorectal coil	Yes in 50 cases; no in 50 cases	No
Sequence	Axial T2-weighted imaging	Axial and sagittal T2-weighted imaging
Slice thickness (mm)	2.2–4.0	3.0
Voxel size (mm)	$0.25-0.75 \times 0.25-0.75$	$0.4-0.7 \times 0.4-0.7$

[a]Twenty-five cases from Haukeland University Hospital, Bergen, Norway; 25 from Beth Israel Deaconess Medical Center, Boston, MA; 25 from University College London, London, UK; 25 from Radboud University Nijmegen Medical Center, Nijmegen, The Netherlands.
[b]All from Korea University Anam Hospital, Seoul, Korea.

ours is the first comparative study of volume determination by CNN-based automatic prostate segmentation versus volume estimation based on the ellipsoid formula.

Our study had limitations. First, we did not use the volume of prostatectomy specimen as our ground-truth volume because not all the patients in our study underwent radical prostatectomy. However, prostate volume measurement by means of planimetry on MR images has exhibited greater precision over other techniques [37]. Second, nearly one-third of cases in our dataset were excluded in the training for segmentation of the transition zone. This caused a discrepancy between the numbers of training datasets for each zone. Third, we used small sample sizes of volume data with large values in our training datasets. Fourth, our proposed CNN has not been independently verified in prior datasets, such as Promise12, and is not currently available for external use. Promise12 data were not included in the evaluation of the performance of our proposed algorithm because the dataset includes only transverse T2-weighted MR images and sagittal images are not available. To measure the volume estimates calculated with the ellipsoid formula, sagittal T2-weighted MR images are necessary to measure the height of the prostate. Therefore, comparison of the volume estimates from our 3D CNN with those of the ellipsoid formula could not be performed with the Promise12 dataset.

Conclusion

Fully automated 3D segmentation with KUProstate provided reliable volume estimates that were superior to the outcomes with the conventional volume measurement model with the ellipsoid formula in segmentation of the entire prostate. Our proposed 3D CNN for fast and accurate volume measurement may help clinicians evaluate prostate diseases properly.

Acknowledgment

References

1. Roobol MJ, Schröder FH, Hugosson J, et al. Importance of prostate volume in the European Randomised Study of Screening for Prostate Cancer (ERSPC) risk calculators: results from the prostate biopsy collaborative group. *World J Urol* 2012; 30:149–155

2. Roobol MJ, van Vugt HA, Loeb S, et al. Prediction of prostate cancer risk: the role of prostate volume and digital rectal examination in the ERSPC risk calculators. *Eur Urol* 2012; 61:577–583

3. Pierorazio PM, Kinnaman MD, Wosnitzer MS, Benson MC, McKiernan JM, Goluboff ET. Prostate volume and pathologic prostate cancer outcomes after radical prostatectomy. *Urology* 2007; 70:696–701

4. Turkbey B, Fotin SV, Huang RJ, et al. Fully automated prostate segmentation on MRI: comparison with manual segmentation methods and specimen volumes. *AJR* 2013; 201:[web]W720–W729

5. Bezinque A, Moriarity A, Farrell C, Peabody H, Noyes SL, Lane BR. Determination of prostate volume: a comparison of contemporary methods. *Acad Radiol* 2018; 25:1582–1587

6. Toth R, Bloch BN, Genega EM, et al. Accurate prostate volume estimation using multifeature active shape models on T2-weighted MRI. *Acad Radiol* 2011; 18:745–754

7. Paterson NR, Lavallée LT, Nguyen LN, et al. Prostate volume estimations using magnetic resonance imaging and transrectal ultrasound compared to radical prostatectomy specimens. *Can Urol Assoc J* 2016; 10:264–268

8. Lee JS, Chung BH. Transrectal ultrasound versus magnetic resonance imaging in the estimation of prostate volume as compared with radical prostatectomy specimens. *Urol Int* 2007; 78:323–327

9. Garvey B, Türkbey B, Truong H, Bernardo M, Periaswamy S, Choyke PL. Clinical value of prostate segmentation and volume determination on MRI in benign prostatic hyperplasia. *Diagn Interv Radiol* 2014; 20:229–233

10. Litjens G, Toth R, van de Ven W, et al. Evaluation of prostate segmentation algorithms for MRI: the Promise12 challenge. *Med Image Anal* 2014; 18:359–373

11. Jia H, Xia Y, Song Y, et al. 3D APA-Net: 3D adversarial pyramid anisotropic convolutional network for prostate segmentation in MR Images. *IEEE Trans Med Imaging* 2019 Jul 11 [Epub ahead of print]

12. Yan K, Wang X, Kim J, Khadra M, Fulham M, Feng D. A propagation-DNN: deep combination learning of multi-level features for MR prostate segmentation. *Comput Methods Programs Biomed* 2019; 170:11–21

13. Zhu Q, Du B, Yan P. Boundary-weighted domain adaptive neural network for prostate MR image segmentation. *IEEE Trans Med Imaging* 2019 Aug 13 [Epub ahead of print]

14. Milletari F, Navab N, Ahmadi SA. V-net: fully convolutional neural networks for volumetric medical image segmentation. In: *Proceedings of the fourth international conference on 3D vision*. Piscataway, NJ: IEEE, 2016; 565–571

15. Mun J, Jang W, Sung DJ, Kim CS. Comparison of objective functions in CNN-based prostate magnetic resonance image segmentation. In: *Proceedings of the IEEE international conference on*

image processing (ICIP). Piscataway, NJ: IEEE, 2017:3859–3863

16. Tewari A, Indudhara R, Shinohara K, et al. Comparison of transrectal ultrasound prostatic volume estimation with magnetic resonance imaging volume estimation and surgical specimen weight in patients with benign prostatic hyperplasia. *J Clin Ultrasound* 1996; 24:169–174

17. Noh H, Hong S, Han B. Learning deconvolution network for semantic segmentation. In: *Proceedings of the IEEE international conference on computer vision (ICCV).* Piscataway, NJ: IEEE, 2015:1520–1528

18. He K, Zhang X, Ren S, Sun J. Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. In: *Proceedings of the IEEE international conference on computer vision (ICCV).* Piscataway, NJ: IEEE, 2015:1026–1034

19. Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. In: Bach F, Blei D, eds. *Proceedings of the 32nd international conference on machine learning.* New York, NY: Association for Computing Machinery, 2015:448–456

20. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR).* Piscataway, NJ: IEEE, 2016:770–778

21. Yu L, Yang X, Chen H, Qin J, Heng PA. Volumetric ConvNets with mixed residual connections for automated prostate segmentation from 3D MR images. In: proceedings of the thirty-first AAAI conference on artificial intelligence. Palo Alto, CA: AAAI Press, 2017:66–72

22. Kingma DP, Ba LJ. Adam: a method for stochastic optimization. In: *Proceedings of the 3rd international conference on learning representations, ICLR 2015.* arxiv.org/abs/1412.6980. Published January 30, 2017. Accessed January 21, 2020

23. Abadi M, Barham P, Chen J, et al. TensorFlow: a system for large-scale machine learning. In: *Proceedings of the 12th USENIX symposium on operating systems design and implementation USENIX symposium on OSDI.* Berkeley, CA: USENIX, 2016:265–283

24. Ma L, Guo R, Zhang G, et al. Automatic segmentation of the prostate on CT images using deep learning and multi-atlas fusion. *Proc SPIE Int Soc Opt Eng* 2017; Feb:10133

25. Mortensen MA, Borrelli P, Poulsen MH, et al. Artificial intelligence–based versus manual assessment of prostate cancer in the prostate gland: a method comparison study. *Clin Physiol Funct Imaging* 2019; 39:399–406

26. Cuocolo R, Cipullo MB, Stanzione A, et al. Machine learning applications in prostate cancer magnetic resonance imaging. *Eur Radiol Exp* 2019; 3:35

27. Aldoj N, Lukas S, Dewey M, Penzkofer T. Semi-automatic classification of prostate cancer on multi-parametric MR imaging using a multi-channel 3D convolutional neural network. *Eur Radiol* 2019 Aug 29 [Epub ahead of print]

28. Guo Y, Gao Y, Shen D. Deformable MR prostate segmentation via deep feature learning and sparse patch matching. *IEEE Trans Med Imaging* 2016; 35:1077–1089

29. Cheng R, Roth HR, Lay N, et al. Automatic magnetic resonance prostate segmentation by deep learning with holistically nested networks. *J Med Imaging (Bellingham)* 2017; 4:041302

30. To MNN, Vu DQ, Turkbey B, Choyke PL, Kwak JT. Deep dense multi-path neural network for prostate segmentation in magnetic resonance imaging. *Int J CARS* 2018; 13:1687–1696

31. Cheng R, Lay N, Roth HR, et al. Fully automated prostate whole gland and central gland segmentation on MRI using holistically nested networks with short connections. *J Med Imaging (Bellingham)* 2019; 6:024007

32. Collette L, Burzykowski T, Schröder FH. Prostate-specific antigen (PSA) alone is not an appropriate surrogate marker of long-term therapeutic benefit in prostate cancer trials. *Eur J Cancer* 2006; 42:1344–1350

33. Sim KC, Sung DJ, Kang KW, et al. Magnetic resonance imaging-based prostate-specific antigen density for prediction of Gleason score upgrade in patients with low-risk prostate cancer on initial biopsy. *J Comput Assist Tomogr* 2017; 41:731–736

34. Loeb S, Bruinsma SM, Nicholson J, et al. Active surveillance for prostate cancer: a systematic review of clinicopathologic variables and biomarkers for risk stratification. *Eur Urol* 2015; 67:619–626

35. Choi YJ, Kim JK, Kim HJ, Cho KS. Interobserver variability of transrectal ultrasound for prostate volume measurement according to volume and observer experience. *AJR* 2009; 192:444–449

36. Karademir I, Shen D, Peng Y, et al. Prostate volumes derived from MRI and volume-adjusted serum prostate-specific antigen: correlation with Gleason score of prostate cancer. *AJR* 2013; 201:1041–1048

37. Jeong CW, Park HK, Hong SK, Byun SS, Lee HJ, Lee SE. Comparison of prostate volume measured by transrectal ultrasonography and MRI with the actual prostate volume measured after radical prostatectomy. *Urol Int* 2008; 81:179–185

38. Jia G, Baudendistel KT, von Tengg-Kobligk H, et al. Assessing prostate volume by magnetic resonance imaging: a comparison of different measurement approaches for organ volume analysis. *Invest Radiol* 2005; 40:243–248

39. Bulman JC, Toth R, Patel AD, et al. Automated computer-derived prostate volumes from MR imaging data: comparison with radiologist-derived MR imaging and pathologic specimen volumes. *Radiology* 2012; 262:144–151

40. Tian Z, Liu L, Zhang Z, Fei B. PSNet: prostate segmentation on MRI based on a convolutional neural network. *J Med Imaging (Bellingham)* 2018; 5:021208

41. Tian Z, Liu L, Fei B. Deep convolutional neural network for prostate MR segmentation. *Int J Comput Assist Radiol Surg* 2018; 13:1687–1696

42. Cheng R, Lay N, Mertan F, et al. Deep learning with orthogonal volumetric HED segmentation and 3D surface reconstruction model of prostate MRI. In: *Proceedings of the IEEE international symposium on biomedical imaging (ISBI).* Piscataway, NJ: IEEE, 2017:749–753