IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON DATA-ENABLED INTELLIGENCE FOR DIGITAL HEALTH

Received August 8, 2019, accepted August 26, 2019, date of publication September 4, 2019, date of current version September 25, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2939389

A Bi-Attention Adversarial Network for Prostate Cancer Segmentation

GUOKAI ZHANG[1], WEIGANG WANG[2], DINGHAO YANG[1], JIHAO LUO[1], PENGCHENG HE[1], YONGTONG WANG[3], YE LUO[1], BINGHUI ZHAO[4], AND JIANWEI LU[1,5]

[1]School of Software Engineering, Tongji University, Shanghai 201804, China
[2]Department of Radiology, Shanghai Fire Corps Hospital, Shanghai 200443, China
[3]College of Electronics and Information Engineering, Tongji University, Shanghai 201804, China
[4]Department of Radiology, Shanghai Tenth People's Hospital, Tongji University School of Medicine, Shanghai 200072, China
[5]Institute of Translational Medicine, Tongji University, Shanghai 200092, China

Corresponding authors: Ye Luo (yeluo@tongji.edu.cn), Binghui Zhao (binghuizhao@163.com), and Jianwei Lu (jwlu33@tongji.edu.cn)

This work was supported in part by the General Program of National Natural Science Foundation of China (NSFC) under Grant 61806147, in part by the Fund of the Science and Technology Commission of Shanghai Municipality, China, under Grant 16411969100, in part by the Shanghai Natural Science Foundation of China under Grant 18ZR1441200, in part by the Fundamental Research Funds for the Central Universities under Grant 22120180012, and in part by NSFC under Grant 81571347 and Grant 61572362.

ABSTRACT Prostate cancer is one of the most prevalent cancers among men. Early detection of this cancer could effectively increase the survival rate of the patient. In this paper, we propose a Bi-attention adversarial network for the prostate cancer segmentation automatically. The proposed architecture consists of the generator network and discriminator network. The generator network aims to generate the predicted mask of the input image, while the discriminator network aims to further improve the generator performance with adversarial learning by discriminating the generator predicted mask and the true label mask. For better improving the segmentation performance, we combine two attention mechanisms with the generator network to learn more global and local features. Extensive experiments on the T2-weighted (T2W) images have demonstrated our model could achieve state-of-the-art segmentation performance compared with other methods.

INDEX TERMS Prostate cancer segmentation, adversarial learning, attention mechanism, generator network, discriminator network.

I. INTRODUCTION

Prostate cancer becomes one of the most prevalent cancers among men in the United States [1]. The early detection and diagnosis of this cancer could efficiently increase the survival rate of the patient. Currently, the most commonly used for prostate cancer detection is by transrectal ultrasound (TRUS) biopsy, and prostate specific antigen (PSA) blood test. Many trial tests have proved that the PSA and TRUS could efficiently reduce prostate cancer mortality by 20% - 30% [2]. Although the clinical examination of the PSA and TRUS has been widely used, it is still limited by the low specificity and degraded diagnosis. Meanwhile, several studies [3], [4] have proved that the magnetic resonance imaging (MRI) could be a potential modality to improve the diagnosis accuracy with a noninvasive way. However, the prostate cancer diagnosis of MRI usually requires professional and experienced radiologists, and the diagnosis process is usually a laborious and repeated work. Thus, the computer-aided detection (CAD) system provides an alternative approach to help radiologist achieve prostate cancer detection automatically.

Traditionally, the CAD system for prostate cancer detection is usually based on the hand-crafted features. According to the ways that the hand-crafted features are used, these methods can be further categorized into the supervised and the unsupervised methods. The unsupervised methods such as the thresholding [5], region growing [6], edge detection and grouping [7], Markov Random Fields (MRFs) [8], Mumford-Shah functional based frame partition [9], level sets [10], graph cut [11], mean shift [12], and their extensions and integrations [13], [14] usually utilize constraints about image intensity or object appearance. Supervised methods [15]–[18], on the other hand, directly learn from labeled training samples, extract features and context information to perform a dense pixel (or voxel)-wise classification. For example, Chan *et al.* [19] proposed a multi-channel statistical

The associate editor coordinating the review of this manuscript and approving it for publication was Ying Song.

classifier by combining information from three different magnetic resonance methodologies. It combined co-occurrence matrix (CM), support vector machine (SVM) and fisher linear discriminant (FLD) for better prostate cancer detection. Liu *et al.* [20] designed an unsupervised segmentation method for prostate cancer detection based on fuzzy MRFs modeling. In addition, Ozer *et al.* [21] presented a Relevance Vector Machines (RVM) for automatic prostate cancer localization to produce more accurate and efficient segmentation results. Yan *et al.* [22] segmented the prostate by putting forward a method of utilizing a prior shape estimated from partial contours. Although those hand-crafted methods have achieved great success in this field, it still remains the challenges of defining the features manually and subjectively.

Recently, with the remarkable performance of the deep learning methods in computer vision tasks, the analysis of this method in medical image segmentation has also been attempted [24], [25], [28], [29]. Litjens *et al.* [29] presented a fully automated computer-aided detection system in segmenting the candidate regions and obtaining the likelihoods of cancer. Guo *et al.* [28] put forward a deformable segmentation method by unifying deep feature learning with the sparse patch matching. Karimi *et al.* [24] proposed a two-step deep-learning based method using two separated convolutional neural networks (CNNs), and then a novel method [25] by employing statistical shape was implemented to predict the location of the prostate cancer. However, due to the segmentation accuracy limitations of patch based methods [28], region based methods [29], and even with some posterior processions [25], all these methods still can not achieve a satisfactory result to pixel-wise level.

Fully convolutional network (FCN) which uses a whole image as input and predicts a pixel-wise level segmentation result [34] has achieved great success in semantic image segmentation. Ronneberger et al. took the idea of the FCN one step further and presented an framework called U-Net [33], which is a regular CNN followed by an up-sampling operation, where up-convolutions are used to increase the size of feature maps. After that time, U-Net or FCN becomes the popular backbone network of various medical segmentation methods [27], [39]–[41]. Such as in [40] and [41], they extended a two-dimensional (2-D) FCN into a volumetric fully ConvNet (3D-FCN) to enable volume-to-volume segmentation prediction. Milletari *et al.* [39] proposed a 3-D variant of U-Net architecture called V-Net for prostate segmentation. Tian *et al.* [27] designed PSNet, which is a fine tuned FCN trained end-to-end in a single learning stage to solve the voxels-unbalance problem. Despite the fact that U-Net like methods can learn low-level and high-level features, there are two limitations of these methods: (1) Unlike traditional conditional random field (CRF) and graph cut methods which are usually adopted for segmentation refinement by incorporating spatial correlation, there is no guarantee of spatial consistency in the final U-Net segmentation map [43]. (2) These approaches lead to excessive and redundant use of computational resources, and similar low-level features are repeatedly extracted by all models within the network architecture.

In order to solve the above problems in U-Net, current researchers try to improve the performance from the following two aspects. On one hand, adversarial losses as introduced by the discriminator in Generative Adversarial Network (GAN) can take into account high order potentials [42], and the spatial correlation either from low-level or high-level features can be regularized globally. Kohl *et al.* [30] adopted the GAN network architecture in [37] to improve the prostate and the cancer region detection performance and evaluated on the self collected dataset. On the other hand, in order to automatically learn to focus on target structures of varying shapes and sizes, an attention gate (AG) model is applied into U-Net and termed Attention U-Net [36]. However, there are few methods which can solve the aforementioned problems simultaneously. Moreover, the attention gate model used in [36] only weighted the learned features globally, and the features selected from a local view are also key to the accurate segmentation.

To address the above challenges, in this paper, we propose a Bi-attention adversarial network for the prostate cancer segmentation, which can select features from global and local views simultaneously, and the adopted GAN network architecture can further ensure the spatial correlations for all the features in the final segmentation map. Specifically, the main designed architecture consists of the generator network and discriminator network. The generator network aims to generate the predicted mask of the input image and we use U-Net as the backbone. The discriminator network aims to further improve the generator performance with adversarial learning by discriminating the generator predicted mask and the true label mask. In order to weight the regional features different region scopes, we propose two attention mechanisms (i.e. channel attention and position attention) with the generator network to improve the segmentation performance. The channel attention mechanism is the utilization of the previous work [31], which calculates the importance weights of each channel, and then use them to highlight the more useful channel features globally. For the position attention mechanism, we aim at extracting more subtle and pixel-level feature information of the input. Due to the skip connections of the original U-Net provide more location and boundary cues from the earlier layers, thus, in our designed model, we combine the position attention mechanism with the skip connection to further improve the ability to extract the subtle and pixel-level feature information.

The main contributions of the proposed method are two folds:

1) We propose a Bi-attention adversarial network for the prostate cancer segmentation, which achieves competitive segmentation results by combining attention mechanisms with the generator network of GAN. Through this way, important features sharing strong spatial correlations are selected thus good segmentation results are obtained.

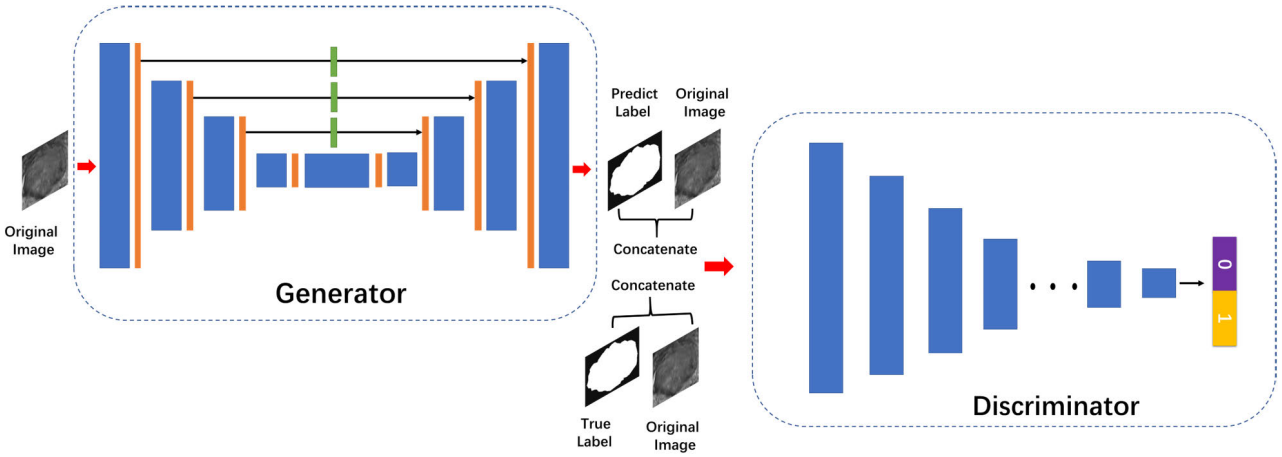


FIGURE 1. The main architecture of the proposed Bi-attention adversarial network for segmenting the prostate cancer. The generator is to generate the predicted mask while the discriminator aims to further improve the generator performance by discriminating the predicted mask and the true label mask.

2) We propose to use channel attention and position attention simultaneously in one single network, thus key features to prostate cancer region segmentation are selected globally and locally.

II. PROPOSED METHOD

In this paper, we propose a novel Bi-attention adversarial network for segmenting the prostate cancer. In the following sections, we will illustrate the main modules of our proposed model in details.

A. NETWORK STRUCTURE

The main architecture of our designed network is illustrated in Figure 1, and it consists of the generator and the discriminator network, respectively.

For the generator, it aims to generate the predicted mask of the input image. We use U-Net as the backbone of the generator network, and it is composed of the encoder and the decoder two stages to generate the predicted mask. For the discriminator, it aims to further improve the generator performance with adversarial learning by discriminating the generator predicted mask and the true label mask.

Especially, to emphasize the features which are contributed to the prostate cancer segmentation, we utilize two attention mechanisms to allow the designed network to extract the attention features globally and locally. More details about those designed structures are presented in the following sections.

B. GENERATOR NETWORK

The main backbone of our generator network is U-Net, which has achieved great success in the medical image segmentation task. U-Net is composed of the encoder stage and the decoder stage. At the encoder stage, the high-level contextual information is extracted by using successive convolutions and pooling layers, while the decoder stage upsamples the extracted encoder high-level feature maps to original image

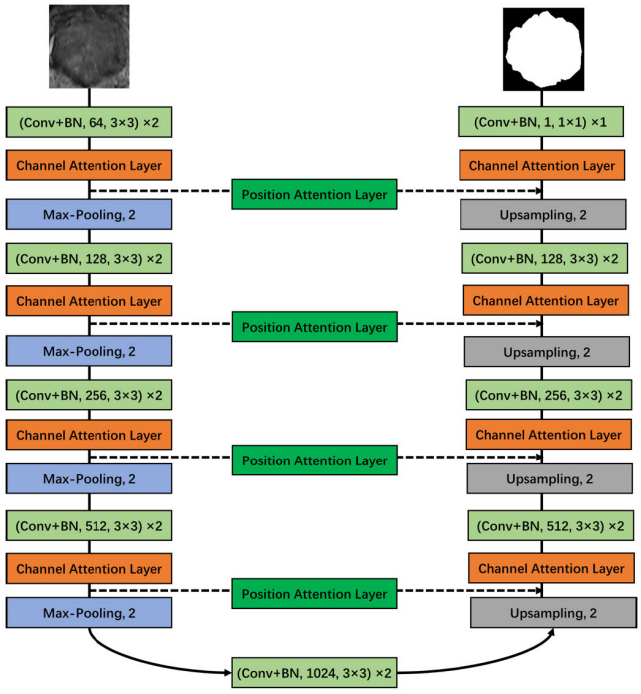


FIGURE 2. The structure of the generator network which including the encoder and the decoder stages. In both encoder and decoder stages, the channel attention layer is embedded after each convolution and batch-normalization layer. And each position attention layer is implemented with the skip connection in the backbone U-Net network.

size to form the final predicted mask gradually. The detailed structure and parameters of the generator network are shown in Figure 2, which adds the channel attention layer after each convolution layer in both encoder and decoder stages, and combines each position attention layer with the skip connection. Here, we adopt four skip connections. With these two specifically designed attention layers, the global information is learned channel-wisely, and the subtle and pixel-level feature information is learned locally. The detailed structure of the channel attention layer and the position attention

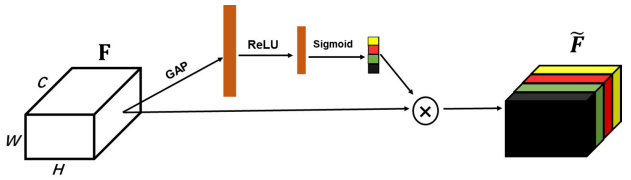


FIGURE 3. The structure of the channel attention layer. The intermediate feature F is first performed with global average pooling (GAP) and then information is further aggregated through ReLU and Sigmoid activation consequently. The obtained channel attention is used to weight F aiming to select channel-wise important features globally as $\widetilde{F}$.

layer are introduced in Section II-B.1 and Section II-B.2, respectively.

1) CHANNEL ATTENTION LAYER

Inspired by [31], which calculates the importance weights of each channel, and then uses them to highlight the more useful channel features. We implement our channel attention layers in generator network such that the inter-dependencies between the channels of its convolutional features are explicitly modeled by the channel attention layer. The detailed structure of the proposed channel attention layer can be found in Figure 3.

Denote the intermediate feature $F \in \mathbb{R}^{H\times W\times C}$ (Here, H, W, and C is the height, the width and the channel of F, respectively.) of the generator network as:

$$F = [F_1, F_2, ..., F_i, ..., F_C], \tag{1}$$

where F_i represents the i_{th} channel feature map of F, $i \in \{1, 2, ..., C\}$. As illustrated in Figure 3, for each F, we first perform a global average pooling (GAP in short) over it to generate channel-wise statistics as $z \in \mathbb{R}^{1\times 1\times C}$ which is the global averaged vector of F. Each z_i is the i_{th} channel statistic which can be calculated as:

$$z_i = \frac{1}{H \times W} \sum_{x}^{H} \sum_{y}^{W} F_i(x, y). \tag{2}$$

After that, to fully capture channel-wise dependencies, a gating mechanism with a sigmoid activation is adopted to achieve information aggregation as:

$$z' = \sigma\left(W_2\delta(W_1 z)\right), \tag{3}$$

where δ is the ReLU activation and σ is the sigmoid activation function. $W_1 \in \mathbb{R}^{C\times \frac{C}{r}}$ and $W_2 \in \mathbb{R}^{\frac{C}{r}\times C}$ represent the weights of the two fully connected layers, respectively. We set $r = 4$ as the bottleneck to reduce the dimension. Here, the ReLU activation is used to ensure that multiple channels can be emphasized while the sigmoid activation is employed to model the nonlinear relationships among different channels. By this way, the importance of each feature channel can be learned and described by z'.

Finally, the channel attention features $\widetilde{F}$ can be obtained by multiplying F with z' channel-wisely as:

$$\widetilde{F} = F * z' = [F_1 * z'_1, F_2 * z'_2, ..., F_i * z'_i, .., F_C * z'_C]. \tag{4}$$

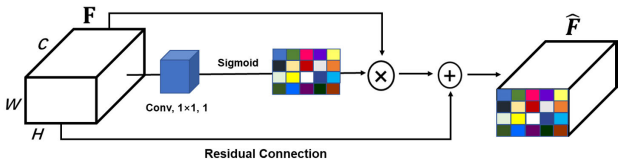


FIGURE 4. The structure of the position attention layer. The intermediate feature F is first convoluted with 1 × 1 × 1 kernel then activated by a Sigmoid to form position attention for all pixels. After weighted with position attention, the position attention feature $\widehat{F}$ is obtained with a residual connection of F.

Through the channel attention, the designed generator network can learn more global information by selectively strengthening informative features and suppressing less useful ones. Therefore, an effective global representation of the input is obtained, and the performance of the segmentation model is expected to be further improved by adding the channel attention layers in both the encoder and the decoder stages.

2) POSITION ATTENTION LAYER

In original U-Net, the skip connections between the encoder and the decoder stages can provide the location and boundary cues from the earlier layers. However, in medical image segmentation, local information plays a vital role to obtain an accurate segmentation result, thus, in our designed model, we propose the position attention and implement with the skip connection to further improve the ability to extract the subtle and pixel-level feature information. The detailed structure of the proposed position attention layer can be found in Figure 4.

Here, we consider the input feature as:

$$F = [F_{1,1}, F_{1,2}, ..., F_{h,w}, ..., F_{H,W}], \tag{5}$$

where $F_{h,w} \in \mathbb{R}^{1\times 1\times C}$ denotes the feature at the position (h, w) on feature map F.

During the position attention operation, we first apply a convolution with a size of $1 \times 1 \times 1$ to F. By this way, we fuse features of all the channels as $F_p \in \mathbb{R}^{H\times W}$. To this end, the attention is focused on the feature position selection, and the features contributed most to the segmentation performance are expect to be emphasized spatially. A sigmoid activation σ is adopted to F_p to form the weight matrix W_p of all pixel positions.

$$W_p = \frac{e^{F_p}}{1 + e^{F_p}} \tag{6}$$

After that, we perform a element-wise multiplication between F and W_p, meanwhile, a residual connection is utilized to gain the final position attention features $\widehat{F}$.

$$\begin{aligned}\widehat{F} &= F \cdot W_p + F \\ &= [F_{1,1} \cdot w_{p(1,1)}, F_{1,2} \cdot w_{p(1,2)}, ..., F_{H,W} \cdot w_{p(H,W)}] + F,\end{aligned} \tag{7}$$

C. DISCRIMINATOR NETWORK

The main structure of our designed discriminator network is illustrated in Figure 5. It contains seven convolution layers

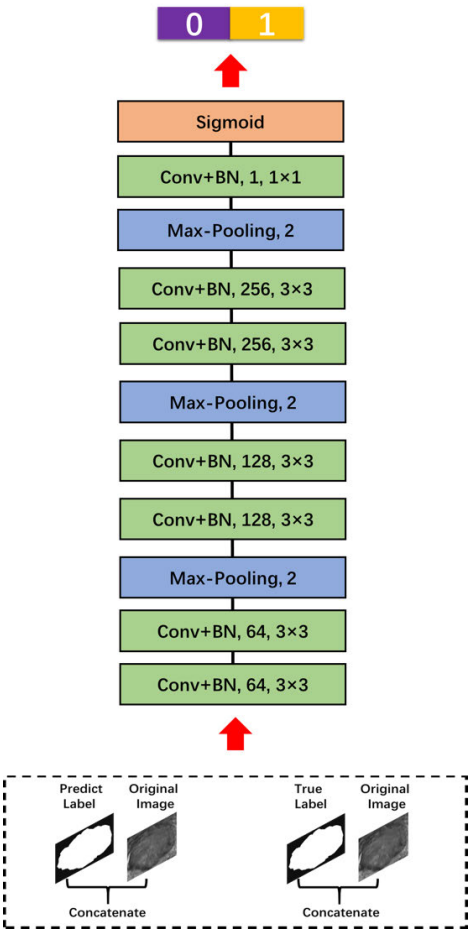


FIGURE 5. The structure of the discriminator network, which contains seven convolution layers and three max-pooling layers. In order to discriminate between the predicted mask and the ground truth mask, and also utilize the information of the original image, we concatenate the original image with the predicted mask and the ground truth mask, respectively. And the two concatenated images are the input of the discriminator network.

and three max-pooling layers. Especially, the LeakyReLU is used as the activation during the training. In order to discriminate between the predicted mask and the ground truth mask, and also keep the information from the original image, we concatenate the original image with the predicted mask and the ground truth mask, respectively. The two groups of concatenated images are as input and then fed into successive convolution and pooling layers to extract high-level features, and the final layer with a sigmoid activation is utilized to predict the binary label. If the predicted score is above 0.5, then the discriminator network regards those inputs as real (1), otherwise as fake (0).

D. LOSS FUNCTIONS

During the network training process, the aim of the generator network G is to generate the similar data based on the true label mask, while the aim of the discriminator network D is to distinguish the generated ones and the real data as possible as it can. Thus, the final training loss could be regarded as a min-max game, and it can be formulated as:

$$L_{final} = \mathbb{E}_{x,y\sim p_{data}(x,y)}[logD(x,y)] + \mathbb{E}_{x\sim p_{data}(x)}[log(1-D(x,G(x)))], \quad (8)$$

where x denotes the original image and y represents the true label mask. Notably, we combine x with $G(x)$ to train the network as the conditional GAN [32].

1) Generator Loss: The loss L_G for the generator network is a binary cross-entropy loss, and it can be formulated as:

$$L_G = -\frac{1}{N}\sum_{i=1}^{N} p_i log(t_i) + (1-p_i)log(1-t_i), \quad (9)$$

where p_i is the predicted label map, and t_i represents the true label map of the original image.

2) Discriminator Loss: The loss L_D for the discriminator network is similar to the generator network by using the binary cross-entropy loss.

Finally, we perform addition operation of L_G and L_S to gain the final loss of the network:

$$L_{final} = L_G + \lambda L_D, \quad (10)$$

where λ is the weight parameter to balance those two losses to be comparable. Discussions about how this parameter affects the final segmentation performance are provided at the experimental Section III-G.

III. EXPERIMENTAL RESULTS AND ANALYSIS

A. DATA INTRODUCTION

The experimental data consist of 120 patients from the Shanghai Tenth People Hospital, Tongji University and we use the images of MR T2-weighted with fat saturation as our experiment modality. The voxel size is $0.9 \times 0.6 \times 3.5mm^3$ of the T2-weighted image. The images were acquired with 3.0 Tesla (T) whole-body unit MR imaging system 9Magnetom Verio 3.0T, Simens Medical Company). All the prostate cancer annotations are confirmed by the biopsy pathology. For better segmentation performance, we first crop the region of interest based on the connected component analysis. Then, we resize all the cropped images and masks to 256×256 as the input to the designed network.

B. IMPLEMENTATION DETAILS

The designed model is implemented based on Tensorflow deep learning library on a workstation with NVIDIA GTX 1080 GPU. The initial learning rate is 10×10^{-4}, and it decays by 0.0005 after 20 epochs. The parameters of the network are initialized by Gaussian Distribution, and we use Adam optimizer to optimize the network. The input size of the original image is 256×256. We use random flip, rotation, and cropping to augment the training data. The network training is conducted by an alternating fashion: the generator network first generates the predicted masks by using the mini-batch data, and then we feed the predicted masks into the discriminator network to update the parameters. After that,

TABLE 1. The effectiveness of different attention modules in our model.

Method	AC (%)	SE (%)	JA (%)	DI (%)
N-ChlPsnAtt	85.2	91.6	73.2	84.1
ChlAtt	86.3	92.8	74.2	85.3
PsnAtt	86.0	92.5	74.7	85.7
Ours	**86.4**	**93.2**	**75.7**	**85.9**

the generator network is updated by training another new mini-bath data.

C. EVALUATION METRICS

In this paper, we use accuracy (AC), sensitivity (SE), jaccard index (JA), and dice coefficient (DI) as the basic metrics to evaluate our designed model. Denote TP the true positive, FP the false positive, TN the true negative, and FN the false negative.

$$AC = \frac{TP + TN}{TP + FP + TN + FN} \tag{11}$$

$$SE = \frac{TP}{TP + FN} \tag{12}$$

$$JA = \frac{TP}{TP + FP + FN} \tag{13}$$

$$DI = \frac{2 * TP}{2 * TP + FP + FN} \tag{14}$$

D. THE EFFECTIVENESS OF DIFFERENT ATTENTION MODULES

In order to show the effectiveness of different attention modules in our model, we perform the following experiments. We compare our proposed model with our generator without the channel attention layer and position attention layer (N-ChlPsnAtt), the generator network only with the channel attention layer (ChlAtt), and the generator network only with the position attention layer (PsnAtt). The detailed comparison results can be seen in Table 1. From this table, we can see, with the single designed attention mechanism, the performance of the model could be improved compared with the N-ChlPsnAtt. Moreover, with channel attention module only, our method achieves higher AC and SE values, but lower JA and DI values. That further validates that the proposed channel attention module and the position attention module are complementary to each other. Hence, the best performance is gained by using the two attention mechanisms together. By employing those two attention modules, our model could learn more global and local attention representations which further improve the segmentation performance of the proposed model.

Detailed visualization examples of different attention modules are shown in Figure 6. From this figure, we can see that the segmented masks of our method are more close to the ground truth compared to N-ChlPsn. Moreover, our PsnAtt method tends to detect more detailed boundaries, while our ChlAtt method puts their attention on the global structure of the cancer region.

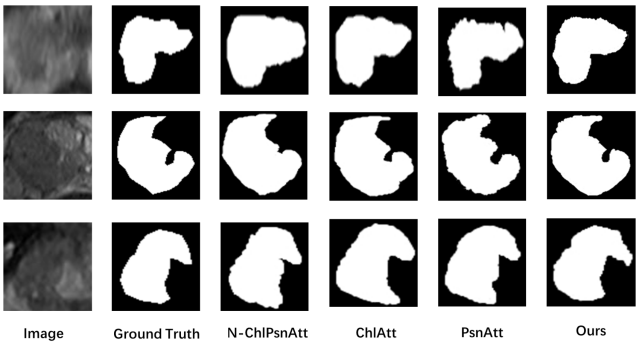


FIGURE 6. The visualization some segmentation results of different attention modules of our method.

TABLE 2. The effectiveness of our channel attention at the encoder and the decoder stages of the generator respectively.

Method	AC (%)	SE (%)	JA (%)	DI (%)
N-ChlPsnAtt	85.2	91.6	73.2	84.1
E-ChlAtt	85.8	92.7	74.0	84.9
D-ChlAtt	85.5	92.6	73.6	84.6
ChlAtt	**86.3**	**92.8**	**74.2**	**85.3**

TABLE 3. Comparisons with position attention layers at different skip connection combinations of U-Net.

Method	AC (%)	SE (%)	JA (%)	DI (%)
N-ChlPsnAtt	85.2	91.6	73.2	84.1
PsnAtt-1	85.3	91.3	73.4	84.3
PsnAtt-2	85.4	91.7	73.0	84.5
PsnAtt-3	85.2	91.9	73.5	84.8
PsnAtt-4	85.6	92.1	74.0	84.6
PsnAtt-12	85.7	91.7	73.2	84.4
PsnAtt-13	85.3	91.9	73.7	84.4
PsnAtt-14	85.7	91.9	74.3	84.8
PsnAtt-23	85.3	92.1	74.0	85.0
PsnAtt-24	85.6	92.0	73.9	85.2
PsnAtt-34	85.8	92.3	74.1	85.3
PsnAtt-123	85.8	92.1	73.9	85.0
PsnAtt-124	85.7	92.3	74.1	84.8
PsnAtt-134	85.9	92.2	74.5	85.3
PsnAtt-234	85.8	92.4	74.6	85.6
PsnAtt	**86.0**	**92.5**	**74.7**	**85.7**

E. THE EFFECTIVENESS OF CHANNEL ATTENTION AT THE ENCODER AND THE DECODER STAGES

Then, we compare the performance of the generator with/without the designed channel attention at the encoder and the decoder stages respectively. We name the generator

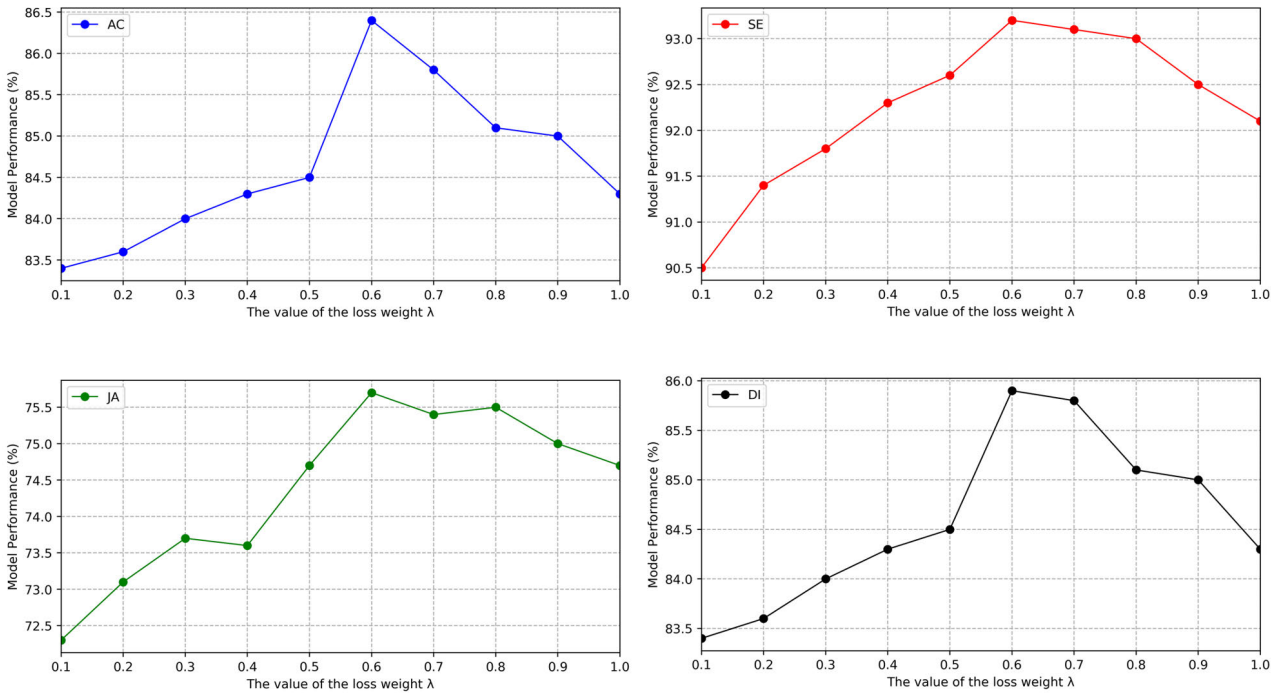


FIGURE 7. The effects of different loss weights with different evaluation metrics: AC on the top left, SE on the top right, JA on the bottom left and DI on the bottom right.

network only with the channel attention layer during the encoder phase as E-ChlAtt and the generator network only with the channel attention layer during the decoder phase as D-ChlAtt. We keep the position attention during these groups of comparisons untouched.

And the comparison results can be seen from Table 2. The result shows that the overall performance of the the E-ChlAtt and the D-ChlAtt could gain better performance than the N-ChlPsnAtt, it further proves the effectiveness of the channel attention mechanism. Meanwhile, it seems that the overall performance of E-ChlAtt is a little higher than the D-ChlAtt. It could be explained that the encoder stage tends to provide more global representations compared with the decoder stage. Overall, the combination of the E-ChlAtt and the D-ChlAtt achieves the best result. Thus, in this paper, we use the channel attention mechanism in both encoder and the decoder stage to improve the final performance.

F. THE COMPARISONS WITH POSITION ATTENTION LAYERS AT DIFFERENT SKIP CONNECTIONS

In this section, we perform the comparisons for the position attention layer at different skip connections. We name the generator network with the position attention layer during the first skip connection as PsnAtt-1, the second skip connection as PsnAtt-2, the third skip connection as PsnAtt-3, and the fourth skip connection as PsnAtt-4. We also performed the comparisons for the position attention layer with different skip connection combinations. The PsnAtt-ij denotes the combination of i_{th} skip connection and j_{th} skip connection,

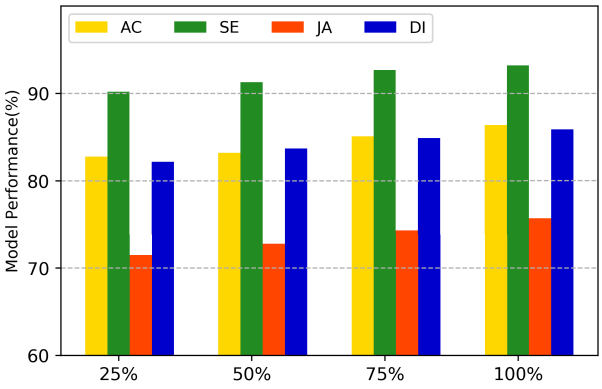


FIGURE 8. Comparison with different data samples. The sizes of the experiment data samples are 25%, 50%, 75%, and 100%, respectively.

and the PsnAtt-ijk represents the combination of i_{th} skip connection, j_{th} skip connection and k_{th} skip connection, where $i, j, k \in \{1, 2, 3, 4\}$.

The detailed comparison results are illustrated in Table 3. For single skip connection, the PsnAtt-3 and the PsnAtt-4 tend to be the most effective ones. It could be explained that the previous layers usually contain more subtle and pixel-level features which further improve the performance of the model. For different skip connection combinations, the best result is achieved by combing the four skip connections simultaneously, while other combinations generally have similar performances. Through this experiment, we can conclude that adding more skip connections could boost the overall performance of the model.

TABLE 4. Comparisons with state-of-the-arts.

Method	AC (%)	SE (%)	JA (%)	DI (%)
U-Net [33]	85.4	92.3	71.3	83.8
FCN [34]	84.9	91.0	72.1	83.2
SegNet [35]	85.9	91.2	72.8	84.1
Attention U-Net [36]	86.1	92.3	72.5	85.3
GAN [37]	85.2	91.6	73.2	84.5
Deeplab V3+ [38]	86.2	**93.5**	73.3	85.2
Ours	**86.4**	93.2	**75.7**	**85.9**

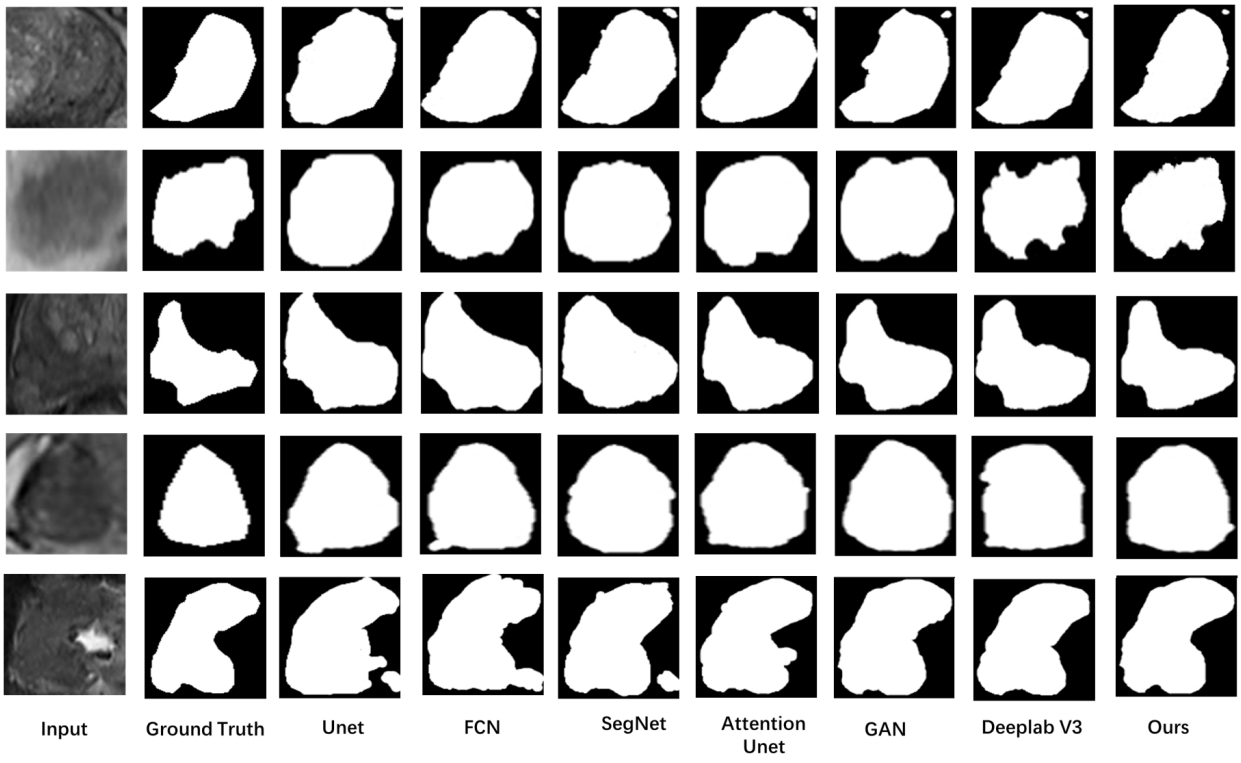


FIGURE 9. Examples of the segmented results for different models.

G. RESULTS OF DIFFERENT LOSS WEIGHTS

In this section, we conduct comparable experiments to explore the effect of different loss weights. Different value of the loss weight λ would give a weighted adjustment of each loss function, and therefore further influence the final performance of the model. The detailed comparison result of different loss weights is depicted in Figure 7. The experimental results with different evaluation metric 9AC, SE, JA and DI) all show that the model increased the performance till the loss weight λ from 0.1 up to 0.6. However, given the heavy weight ($\lambda > 0.6$) of the λ could also have an unfavorable influence on the segmentation performance. Thus, the final value of λ is set to 0.6 empirically for achieving a compromising result.

H. COMPARISON WITH DIFFERENT TRAINING DATA SAMPLES

As illustrated in Figure 8, we analyze the model performance with different training data samples. The sizes of the experiment training data samples are 25%, 50%, 75%, and 100%, respectively. The best result performance is achieved by 100% data samples with 86.4% AC, 93.2% SE, 75.7% JA and 85.9% DI, respectively. The result also shows that with more data samples, the better performance of the model could achieve. That is reasonable due to more data samples could provide more diverse feature representations, which helps the network improve the final segmentation ability. We will try to collect more data to enhance the overall performance in future works.

I. COMPARE WITH STATE-OF-THE-ARTS

To evaluate the performance of our designed model, we compare our method with state-of-the-art segmentation methods. Except for U-Net [33], FCN [34], Attention U-Net [36] and GAN [37], we add two more classical semantic segmentation methods SegNet [35] and Deeplab V3+ [38] for thorough comparisons. For a fair comparison, we re-implement the U-Net, FCN, SegNet, Attention U-Net, GAN, Deeplab V3+.

The detailed comparison result is illustrated in Table 4. The result demonstrates that our proposed model could achieve competitive results especially in AC, JA, and DI. Although SE is not gained the best result, it can be explained that the structure of the Deeplab V3+ method with spatial dilated pooling could learn more multi-scale features. Meanwhile, the comparison between GAN and our proposed model further proves the effectiveness of our designed attention mechanisms.

Figure 9 shows some predicted examples of state-of-the-art models. We compare U-Net, FCN, SegNet, Attention U-Net, GAN, and Deeplab V3+. Compared with the traditional segmentation models (U-Net, FCN, SegNet), our model could gain some obvious improvements. Such as in the last row, extra regions are detected by these methods, while our method successfully filters out these distracts. Compared with Attention U-Net, GAN and Deeplab V3+, our method can achieve a more accurate boundary and shape of the cancer region. For example, in the second and the forth rows, Attention U-Net and GAN models fails to get the exact shapes of the cancer region, and Deeplab V3+ seems to over-fit the cancer region. Only the detection result by our method is close to the ground truth. That could be explained that with our two designed attention mechanisms more global and local subtle representations are provided to further boost the identification of the prostate cancer regions.

IV. CONCLUSION

In this paper, we propose a Bi-attention adversarial network for prostate cancer segmentation. Compared with other state-of-the-art methods, our model could extract more global and local attention features with the channel attention layer and position attention layer by adversarial learning strategy. In the future work, we will try on different MRI modalities to segment the prostate cancer automatically.

ACKNOWLEDGMENT

(Guokai Zhang and Weigang Wang contributed equally to this work.)

REFERENCES

[1] A. Jemal, R. Siegel, E. Ward, T. Murray, J. Xu, and M. J. Thun, "Cancer statistics, 2007," *CA, A Cancer J. Clinicians*, vol. 57, no. 1, pp. 43–66, 2007.

[2] F. H. Schröder *et al.*, "Screening and prostate-cancer mortality in a randomized European study," *New England J. Med.*, vol. 360, no. 13, pp. 1320–1328, Mar. 2009.

[3] H. Hricak, P. L. Choyke, S. C. Eberhardt, S. A. Leibel, and P. T. Scardino, "Imaging prostate cancer: A multidisciplinary perspective," *Radiology*, vol. 243, no. 1, pp. 28–53, 2007.

[4] S. E. Seltzer, D. J. Getty, C. M. Tempany, R. M. Pickett, M. D. Schnall, B. J. McNeil, and J. A. Swets, "Staging prostate cancer with MR imaging: A combined radiologist-computer system," *Radiology*, vol. 202, no. 1, pp. 219–226, 1997.

[5] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-9, no. 1, pp. 62–66, Jan. 1979.

[6] R. Adams and L. Bischof, "Seeded region growing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 16, no. 6, pp. 641–647, Jun. 1994.

[7] J. F. Canny, "A computational approach to edge detection, Readings in computer vision: Issues, problems, principles, and paradigms," *IEEE Trans. Pattern*, to be published.

[8] B. S. Manjunath and R. Chellappa, "Unsupervised texture segmentation using Markov random field models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 5, no. 5, pp. 478–482, May 1991.

[9] D. Mumford and J. Shah, "Optimal approximations by piecewise smooth functions and associated variational problems," *Commun. Pure Appl. Math.*, vol. 42, no. 5, pp. 577–685, 1989.

[10] R. Malladi, J. A. Sethian, and B. C. Vemuri, "Shape modeling with front propagation: A level set approach," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 17, no. 2, pp. 158–175, Feb. 1995.

[11] J. Shi and J. Malik, "Normalized cuts and image segmentation," Departmental Papers (CIS), 2000, p. 107

[12] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 5, pp. 603–619, May 2002.

[13] C.-H. Lee, S. Wang, A. Murtha, M. R. G. Brown, and R. Greiner, "Segmenting brain tumors using pseudo–conditional random fields," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Berlin, Germany: Springer, 2008, pp. 359–366.

[14] A. E. Lefohn, J. E. Cates, R. T. Whitaker, "Interactive, GPU-based level sets for 3D segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Berlin, Germany: Springer, 2003, pp. 564–572.

[15] B. H. Menze *et al.*, "The multimodal brain tumor image segmentation benchmark (BRATS)," *IEEE Trans. Med. Imag.*, vol. 34, no. 10, pp. 1993–2024, Oct. 2015.

[16] D. Cobzas, N. Birkbeck, M. Schmidt, M. Jagersand, and A. Murtha, "3D variational brain tumor segmentation using a high dimensional feature set," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, Oct. 2007, pp. 1–8.

[17] E. Geremia, B. H. Menze, O. Clatz, E. Konukoglu, A. Criminisi, and N. Ayache, "Spatial decision forests for MS lesion segmentation in multi-channel MR images," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Berlin, Germany: Springer, 2010, pp. 111–118.

[18] M. Wels, G. Carneiro, A. Aplas, M. Huber, J. Hornegger, and D. Comaniciu, "A discriminative model-constrained graph cuts approach to fully automated pediatric brain tumor segmentation in 3-D MRI," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Berlin, Germany: Springer, 2008, pp. 67–75.

[19] I. Chan, W. Wells, III, R. V. Mulkern, S. Haker, J. Zhang, K. H. Zou, S. E. Maier, and C. M. C. Tempany, "Detection of prostate cancer by integration of line-scan diffusion, T2-mapping and T2-weighted magnetic resonance imaging; A multichannel statistical classifier," *Med. Phys.*, vol. 30, no. 9, pp. 2390–2398, 2003.

[20] X. Liu, D. L. Langer, M. A. Haider, Y. Yang, M. N. Wernick, and I. S. Yetik, "Prostate cancer segmentation with simultaneous estimation of Markov random field parameters and class," *IEEE Trans. Med. Imag.*, vol. 28, no. 6, pp. 906–915, Jun. 2009.

[21] S. Ozer, M. A. Haider, D. L. Langer, T. H. van der Kwast, A. J. Evans, M. N. Wernick, J. Trachtenberg, and I. S. Yetik, "Prostate cancer localization with multispectral MRI based on relevance vector machines," in *Proc. IEEE Int. Symp. Biomed. Imag., From Nano Macro*, Jun./Jul. 2009, pp. 73–76.

[22] P. Yan, S. Xu, B. Turkbey, and J. Kruecker, "Discrete deformable model guided by partial active shape model for TRUS image segmentation," *IEEE Trans. Biomed. Eng.*, vol. 57, no. 5, pp. 1158–1166, May 2010.

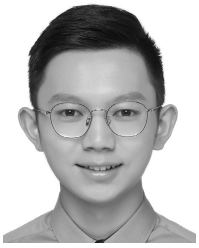
[23] P. Vos, J. O. Barentsz, N. Karssemeijer, and H. J. Huisman, "Automatic computer-aided detection of prostate cancer based on multiparametric magnetic resonance image analysis," *Phys. Med. Biol.*, vol. 57, no. 6, p. 1527, 2012.

[24] D. Karimi, G. Samei, Y. Shao, and T. Salcudean, "A deep learning-based method for prostate segmentation in T2-weighted magnetic resonance imaging," Jan. 2019, *arXiv:1901.09462*. [Online]. Available: https://arxiv.org/abs/1901.09462

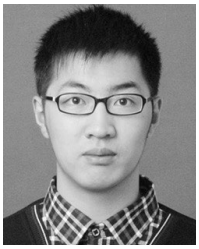
[25] D. Karimi, G. Samei, C. Kesch, G. Nir, and S. E. Salcudean, "Prostate segmentation in MRI using a convolutional neural network architecture and training strategy based on statistical shape models," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 13, no. 8, pp. 1211–1219, Aug. 2018.
[26] Y. Wang, B. Zheng, D. Gao, and J. Wang, "Fully convolutional neural networks for prostate cancer detection using multi-parametric magnetic resonance images: An initial investigation," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, 2018, pp. 3814–3819.
[27] Z. Tian, L. Liu, Z. Zhang, and B. Fei, "PSNet: Prostate segmentation on MRI based on a convolutional neural network," *J. Med. Imag.*, vol. 5, no. 2, 2018, Art. no. 021208.
[28] Y. Guo, Y. Gao, and D. Shen, "Deformable MR prostate segmentation via deep feature learning and sparse patch matching," *IEEE Trans. Med. Imag.*, vol. 35, no. 4, pp. 1077–1089, Apr. 2016.
[29] G. Litjens, O. Debats, J. Barentsz, N. Karssemeijer, and H. Huisman, "Computer-aided detection of prostate cancer in MRI," *IEEE Trans. Med. Imag.*, vol. 33, no. 5, pp. 1083–1092, May 2014.
[30] S. Kohl, D. Bonekamp, H.-P. Schlemmer, K. Yaqubi, M. Hohenfellner, B. Hadaschik, J.-P. Radtke, and K. Maier-Hein, "Adversarial networks for the detection of aggressive prostate cancer," Feb. 2017, *arXiv:1702.08014*. [Online]. Available: https://arxiv.org/abs/1702.08014
[31] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.
[32] M. Mirza and fS. Osindero, "Conditional generative adversarial nets," Nov. 2014, *arXiv:1411.1784*. [Online]. Available: https://arxiv.org/abs/1411.1784
[33] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.
[34] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3431–3440.
[35] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.
[36] O. Oktay, J. Schlemper, L. Le Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert, "Attention U-Net: Learning where to look for the pancreas," Apr. 2018, *arXiv:1804.03999*. [Online]. Available: https://arxiv.org/abs/1804.03999
[37] J. Son, S. J. Park, and K.-H. Jung, "Retinal vessel segmentation in fundoscopic images with generative adversarial networks," Jun. 2017, *arXiv:1706.09318*. [Online]. Available: https://arxiv.org/abs/1706.09318
[38] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 801–818.
[39] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 4th Int. Conf. 3D Vis. (3DV)*, Stanford, CA, USA, 2016, pp. 565–571.
[40] D. Ciresan, A. Giusti, L. M. Gambardella, and J. Schmidhuber, "Deep neural networks segment neuronal membranes in electron microscopy images," in *Proc. 25th Int. Conf. Neural Inf. Process. Syst.*, 2012, pp. 2843–2851.
[41] L. Yu, X. Yang, H. Chen, J. Qin, and P. A. Heng, "Volumetric ConvNets with mixed residual connections for automated prostate segmentation from 3D MR images," in *Proc. 21st AAAI Conf. Artif. Intell.*, 2017, pp. 1–7.
[42] D. Yang, D. Xu, S. K. Zhou, B. Georgescu, M. Chen, S. Grbic, D. Metaxas, and D. Comaniciu, "Automatic liver segmentation using an adversarial image-to-image network," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*. Cham, Switzerland: Springer, 2017, pp. 507–515.
[43] G. Lin, C. Shen, A. van den Hengel, and I. Reid, "Efficient piecewise training of deep structured models for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3194–3203.

GUOKAI ZHANG is currently pursuing the Ph.D. degree with the College of Software Engineering, Tongji University, Shanghai, China. His current research interests include deep learning, object detection, and medical image analysis.

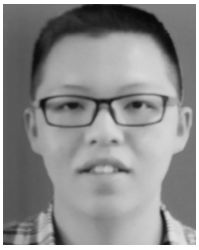
WEIGANG WANG received the M.D. degree from the Medical College, Tongji University, Shanghai, China. He is currently a Chief Physician with the Department of Radiology, Shanghai Fire Corps Hospital, Shanghai. His interests include the prostate cancer diagnosis and medical image analysis.



DINGHAO YANG is currently pursuing the bachelor's degree with the School of Software Engineering, Tongji University. His research interests include deep learning, image and video processing, and computer vision.

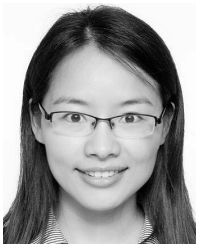


JIHAO LUO is currently pursuing the bachelor's degree in software engineering and projects management with the School of Software Engineering, Tongji University. His research interests include vision science, knowledge graph, and software developing.



PENGCHENG HE is currently pursuing the bachelor's degree in image processing and graphics with the School of Software Engineering, Tongji University. His research interests include vision science and object detection.

YONGTONG WANG received the degree from the College of Electronics and Information Engineering, Tongji University. She is currently a Software Engineer in data mining and system building. Her research interests include deep learning and natural language processing.



YE LUO received the M.S.E. degree in signal and information processing from Anhui University, Hefei, China, in 2008, and the Ph.D. degree from Nanyang Technological University, Singapore, in 2014. She is currently an Assistant Professor with Tongji University. Her research interests include computer vision, machine learning, content-/perceptual-based video analytics, and medical image processing.

BINGHUI ZHAO received the M.D. degree from Medical College, Shanghai Jiaotong University, China, in 2008. He is currently an Associate Chief Physician in charge of abdomen imaging with the Department of Radiology, Shanghai Tenth People's Hospital, Tongji University School of Medicine, Shanghai, China.

JIANWEI LU received the Ph.D. degree from the Department of Computer Science, University of Southern California. He is currently a Professor with the Advanced Institute of Translational Medicine and the School of Software Engineering, Tongji University. His research interest includes vision science.