A Multi-Scale Channel Attention Network for Prostate Segmentation

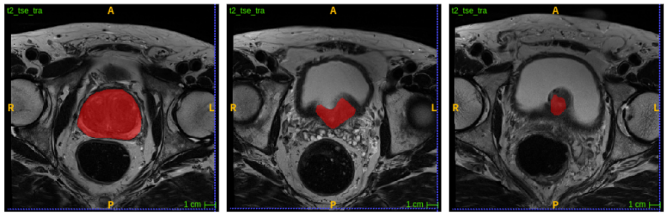
***Abstract*—Prostate cancer is one of the most common malignant tumors in men. Accurate prostate segmentation is essential for prostate cancer diagnosis and intervention. However, the variation in prostate shape, appearance, and size makes the task challenging, given the limit of the annotated data. There are works proposed applying convolutional neural networks in the literature for the task. In this paper, we propose a method using multi-scale and Channel-wise Self-Attention (CSA) to re-calibrate the feature maps from multiple layers. By embedding the multi-scale CSA on the skip-connection in a UNet structure, called as UCAnet, we show the consistent improvement of the prostate segmentation in Dice, IoU and ASSD. For comparison, we also investigate the single-scale CSA in the networks, and incorporate the vision transformer to test if a transformer would boost the performance. Experiments on a public dataset with 204 prostate MRI scans show that UCAnet achieves the best performance and outperforms other state of art methods for prostate segmentation such as ENet, UNet, USE-Net and TransUNet.**

***Index Terms*—Prostate segmentation, CNN, Channel-wise, Multi-scale, Vision Transformer**

I. Introduction

Prostate cancer is one of the most significant health problems in the world that becomes the third most common cancer for humans [1]. The incidence and mortality rates of prostate cancer have been continuously increasing since 2015 [2]. In 2022, it alarmingly accounts for 27% of all the estimated cancer cases diagnosed and 11% of all the estimated cancer deaths for men in the US, having the second greatest number of deaths [3]. Millions of affections occur to men every year [4]. To quantify the possibility of clinically significant cancer, lesions in the gland are scored between 1 to 5 based on the Prostate Imaging Reporting & Data System (PI-RADS) [5].

Prostate segmentation on magnetic resonance images (MRIs) is an important prerequisite to accurate lesion detection and cancer diagnosis. However, the blurry boundaries make manual delineation challenging and time-consuming to distinguish the prostate gland from its surrounding tissue. Moreover, manual segmentation also suffers from subjective criteria and uneven quality among annotators. Consequently, reliable and stable automatic segmentation for the prostate gland comes to play and has received notable attention. At the early stage, machine learning method such as atlas-based and model-based techniques were applied for prostate segmentation. At present, convolutional neural networks (CNNs) have shown its promising ability and become the mainstream method for semantic segmentation, and U-Net is the most popular structure in the medical image domain. Milletari et al. [6] deployed a V-Net on PROMISE12, a public prostate T2-weighted (T2w) MRI dataset [7]. Apart from delineating on 3D patches, Zhu et al. utilized a deeply supervised CNN, which can effectively learn 2D prostate features and pass the features from early layers to later layers [8]. In 2021, three popular Deep Learning (DL) networks, UNet, ENet and ERFNet, were evaluated on prostate segmentation via T2w MRI dataset PROSTATEx [9]. Their results showed that all the networks can segment the prostate with good performance, especially ENet [10]. Apart from CNNs, Cheni et al. [11] utilized a converted spiking neural network (SNN) for image segmentation on a video dataset. Their method achieved a satisfactory accuracy with high convergence speed.



(a) (b) (c)

Fig. 1. Examples of shape and size variance among prostates and among slices from the same prostate: (a) z=15 of sequence 81; (b) z=21 of sequence 81; (c) z=15 of sequence 147.

However, there are still some challenges in this task. Firstly, the shape of prostate varies largely from patient to patient. Within the sequence of one prostate, the shape of 2D slices can also varies a lot. The comparison is shown as Fig.1. Rundo et al. [12] proposed a USE-Net to increase the generalization ability of network. They added a Squeeze-and-Excitation Block [13] after each encoder and decoder. Nowadays, many works aim to capture global information for image segmentation tasks. One of the models using both local and global features is TransUNet [14], which first extracts features by CNN, then feeds them into Transformer layers as the last step of Downsampling.

Secondly, size of gland has enormous difference. To make things worse, a large inter-operator volume difference of prostate can exist. Hamzaoui et al. [15] attempted to solve this problem by using two UFNets. One was for locating and extracting the region of interest, which then served as the input of the second one for segmentation. However, this work is not end-to-end, so it is not efficient enough in practice. In 2022, Wang et al. [16] proposed UCTransNet, which deployed a Channel Transformer module to replace U-Net skip connections, and fused the multi-scale features from the point of channels. Despite that it can obtain satisfactory results,

the involvement of Transformer [17] leads to a huge number of parameters and high computational complexity. In clinical applications, it may lead to high requirements for hardware facilities. Tsai and Huang [18] proposed a Unet-EN for multi-hand segmentation on exhibition in the complex real-world scenes. Based on UNet, they re-encodered the output mask to learn more about features.

To deal with the above problems, we propose an effective and efficient end-to-end network architecture, UCAnet. To further stimulate the capability of skip connections instead of simple concatenation, we uniquely utilize a channel-wise self-attention (CSA) mechanism on skip connections, assigning learnable weights to channels. In this way, the important patterns can be emphasized relative to others. On the other hand, how to acquire multi-scale features and narrow the semantic gap among different layers is another problem. Capturing multi-scale features may be vital for medical image segmentation that has the nature of a complicated scale diversity [16]. Hence, after discarding the spatial distribution information, we utilize a multi-scale light-weighted CSA mechanism to learn relative importance among all channels of different layers.

Our major contributions are two-fold: (1) We deploy a Multi-scale Channel-wise Self-Attention (CSA) mechanism on feature maps from the multiple layers of the encoder to realize the feature recalibration among different scales. (2) We provide a concise and effective CNN-based network UCAnet for prostate segmentation. Based on UNet, it only has an additional Multi-scale CSA block on the skip connection. Our results on PROSTATEx MRI dataset show that UCAnet can be a competitive candidate as the helping tool for the prostate diagnosis.

II. Methods

A. Structure Design

Our UCAnet inherits a traditional 5-layer UNet structure with channel number 64, 128, 256, 512, 1024. It is dedicated to make fully advantage of multi-scale channel-wise features. Fig.2 demonstrates an overview of our network, taking the example of fusing only first three layers for a clearer demonstration. To highlight more semantically useful channels and suppress the noisy ones, a CSA mechanism is applied during skip connections. Second, To further tackle the size variety problems, multi-scale features needs to be captured. Hence, in skip connections, we apply a multi-scale CSA block, and all channels from multiple layers can be weighted according to their semantic importance relative to others. The combination of first four layers had the highest performance.

B. Multi-scale CSA Block

In UCAnet, as presented in Fig.3, the feature maps from multiple layers of the encoder are squeezed by spatial global-average-pooling (GAP) and concatenated together along channel dimension. Let $\mathbf{F}_i \in \mathbb{R}^{C_i \times H_i \times W_i}$ be the feature map having a channel number of C_i from the i-th layer. For example, if $i = 1, 2, 3$, the input $\mathbf{z}$ after GAP and concatenation of these three layers is given by:

$$\mathbf{z} = Concat(GAP(\mathbf{F}_1), GAP(\mathbf{F}_2), GAP(\mathbf{F}_3)) \quad (1)$$

In this way, feature maps of multiple layers are combined into one array $\mathbf{z} \in \mathbb{R}^{(\sum C_i)\times 1\times 1}$. Then, the Rectified Linear Unit (ReLU) [19] activation function and two Fully Connected (FC) layers followed by a Sigmoid activation function σ are used to learn the relative importance of every channel. A series of weights $\mathbf{x} \in \mathbb{R}^{(\sum C_i)\times 1\times 1}$ is derived by:

$$\mathbf{x} = \sigma(\mathbf{W}_2 ReLU(\mathbf{W}_1 \mathbf{z})) \quad (2)$$

Weights $\mathbf{x}$ is partitioned according to channel numbers. Taking again fusing the first three layers as an example, weights $\mathbf{x}_2$ corresponding to the feature map of the second layer is,

$$\mathbf{x}_2 = \mathbf{x}[C_1 : C_2] \quad (3)$$

After splitting the weights for each layer, we expand $\mathbf{x}_i$ to the same dimension as that of $\mathbf{F}_i$ by repeating values spatially. The expanded weight $\mathbf{X}_i \in \mathbb{R}^{C_i \times H_i \times W_i}$ then has a pixel-wise multiplication with $\mathbf{F}_i$ to re-calibrate the channels.

C. Decoder with CSA Block

During the Upsampling process, not all the features focused by certain channels contribute positively to the final performance. Therefore, based on UCAnet, we further recalibrate the features via adding single-scale CSA block after every Upsampling process in decoder and refer this network as UCAnet-D. Single-scale CSA is the simplified version of the multi-scale one. It omits the concatenation and split steps. The outputs from Fully Connected layers will weight corresponding channels of original feature maps directly.

D. Transformer after Encoder Block

We also integrate Transformer into UCAnet and name this network as UCAnet-T. The overview of a Transformer block after the encoder is shown in Fig.4. It has two layers of Transformer blocks. Each block is composed of a 6-head Self-Attention (SA) layer and a Multi-Layer Perceptrons (MLP) layer with Layer Norm added between them. The Block number and head number are carefully chosen through experiments. Feature maps are squeezed into 3 channels first, then they are embedded with a patch size of 16 and flattened. A residual structure is used to avoid vanishing gradient problem. Then, the reconstructed Feature maps are squeezed by GAP and enter the multi-scale CSA block.

III. Experiments and Results

A. Dataset

The model was developed and tested on a public dataset PROSTATEx [9], containing 204 T2w MR subjects. Images were obtained by a turbo spin echo sequence and possessed an approximate resolution of $0.5 \times 0.5 \times 3.6$ mm.

To test the generalization, a test set containing 39 samples was randomly chosen. All benchmarks were performed with a

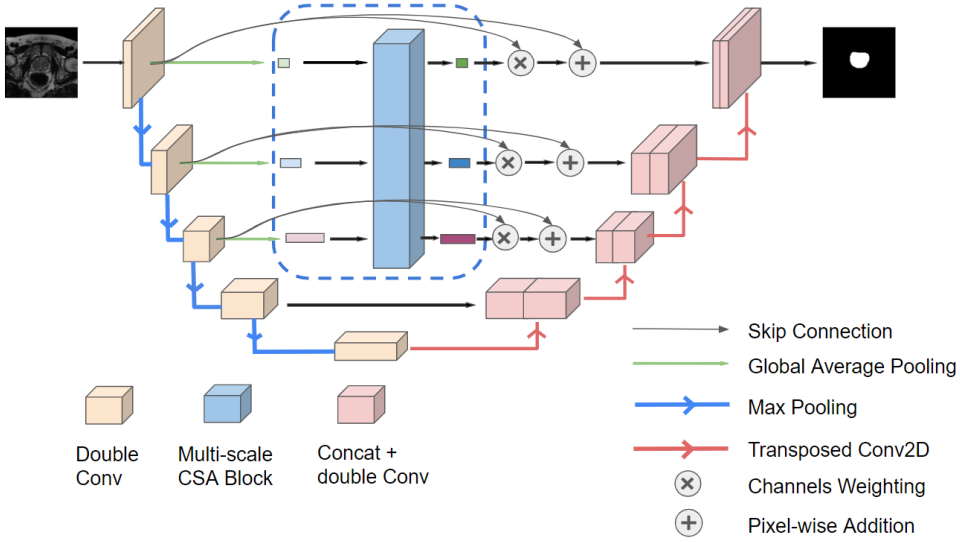


Fig. 2. Structure of the proposed UCAnet.

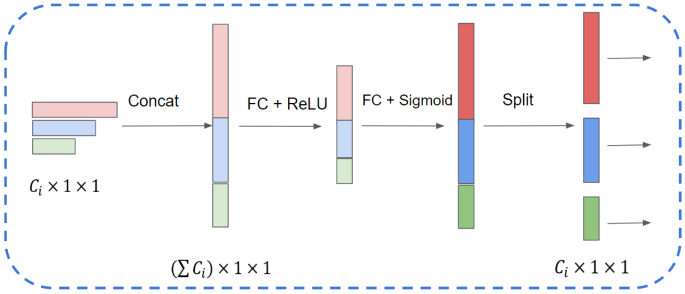


Fig. 3. Multi-scale CSA Block.

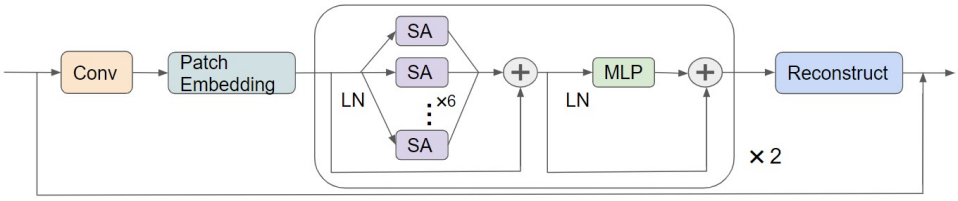


Fig. 4. Transformer block acting on feature maps in UCAnet-T.

5-fold cross-validation. One fold served as validation set, while remaining 4 folds served as training data. 2D slices from the same patient were always split into the same fold for either training, validation or testing. Each image was resized to 256×256 pixels. Intensity values were clipped at a maximum of 900, followed by a linear normalized to $0 - 1$.

Hamzaoui et al. [15] claimed that there were three mismatched data sequences in PROSTATEx dataset. However, through our investigations and feedback from the doctor, all sequences seem satisfactory.

B. Implementation Details

The training of UCAnet was implemented using NVIDIA GeForce or NVIDIA RTX A5000 GPU. It could also be trained on other GPUs with lower memory. We implemented the method on Pytorch version 1.10.2 in Python 3.6 environment.

The training configurations were set as follows: a batch size of 16, an Adam optimizer with an initial learning rate of 5×10^{-4}. All the investigated networks were trained using the Dice Loss [6] to cope with the imbalance between the amount of foreground and background pixels. When validation accuracy did not improve for 5 continuous epochs, we reduced the learning rate by multiplying 0.5. After reduction, we waited for two more epochs before counting the stagnation of validation accuracy again. The model converged if validation accuracy achieved no improvement for 16 epochs. To deal with the problem of limited labelled data, data augmentation including affine transform, Shift-scale-rotate and Gauss Noise

TABLE I
PERFORMANCE RESULTS ON PROSTATEX DATASET.

Networks	DSC (%)	IoU	HD95%	ASSD
ENet [10]	89.91±0.248	81.77±0.390	2.10±0.029	0.682±0.021
UNet [10]	90.79±0.077	83.24±0.114	**1.90±0.050**	0.637±0.031
USE-Net [12]	90.79±0.175	83.23±0.290	1.93±0.129	0.620±0.021
TransUNet [14]	90.79±0.100	83.23±0.164	1.93±0.068	0.616±0.011
UCAnet-T	90.70±0.113	83.07±0.189	1.94±0.039	0.619±0.010
UCAnet-D	90.83±0.223	83.29±0.372	1.95±0.058	0.621±0.019
UCAnet	**90.93±0.092**	**83.45±0.151**	1.99±0.170	**0.614±0.013**

Note: Best results for each metric are in bold. UCAnet recalibrates the feature maps of the first four layers.

was applied to extend training dataset and avoid overfitting. Augmented images were added back into original training pool instead of replacement.

C. Results and analysis

Experimental results on PROSTATEx dataset are shown in Table.I. The Dice Score Coefficient (DSC), Hausdorff distance (95%) (HD95), Intersection over Union (IoU) and average symmetric surface distance (ASSD) were calculated as metrics. They were calculated using the model with highest DSC. We did not include the results of [6], as it did not use the same dataset. Their performance on PROMISE12 achieved 87% DSC. The results show that our method achieves superior or similar results throughout these metrics.

As can be seen from Table.I, our method UCAnet surpasses the classic CNN methods such as UNet, ENet and USE-Net. In the study done by Cuocolo et al. [10], ENet reached a DSC of 90.6%, while UNet had a DSC of 88.1%. We re-implemented the ENet, making configurations the same as mentioned in their paper and fine-tuned the parameters. If we used the same image size 256×256 as all the other models, the best result of ENet was only 89.9% DSC. If inputs were resized to 512×512, its performance could increase to 90.80%. Due to the limitation of hardware, we still stuck to the image size of 256×256 to evaluate all the networks.

The study done by Hamzaoui et al [15] also targeted on PROSTATEx dataset. We did not find the code for their model, UFNet, so we could not re-implement it. Based on their paper, the best single model performance of UFNet was 90.6%. Combining the outputs of five networks resulting from cross-validation, they could obtain a DSC of 90.9%. Single model performance of our UCAnet could reach 90.93%, surpassing the state-of-the-art result. Moreover, using the same training-testing splitting technique, our result had much smaller standard deviation than that of UFNet, which means our method is more stable.

Transformer-based model has been increasingly popular. As mentioned before, TransUNet [14] utilized Transformer to extract global information. However in our testing, it could only reach a similar performance to UNet, while requiring a large number of parameters and long time of training. Our method exceeds TransUNet in terms of higher performance and shorter convergence time.

In addition, we have applied Transformer on feature maps out from multiple layers of encoder. Then, reconstructed feature maps with global information were fed into multi-scale CSA block. This network is called UCAnet-T. The best result obtained was 90.7%. Additionally, to further boost the important features, we have also added a single-scale CSA block after every Upsampling operation of our UCAnet and named it as UCAnet-D. The performance was slightly better at 90.83%. The involvement of more feature recalibration during decoding brings negative effect.

Considering the above experiments we have made, the reason that our UCAnet outperforms other models may have two aspects. First, some global features such as shape and texture vary greatly among sequences [7], [20]. Using the training data may not be able to capture the representative distribution of the testing set. Second, due to the relatively small dataset, models with a mass of parameters may be difficult to adjust parameters to the optima and end up Overfitting.

IV. CONCLUSION

In this work, we propose a concise and effective network UCAnet. The Multi-scale CSA block highlights useful channels considering multiple layers of features. On PROSTATEx dataset, our model achieved an accurate result of 90.93% DSC. Incorporating Vision Transformer to capture the cross-patch attention shows some improvement over some other state-of-the-art works, but does not perform better than the UCAnet, might due to the limitation of the data available for training.

As for future study, we plan to explore the effect of global features and the possibility of a valid combination of Transformer and CNN-based model with more unlabelled data for prostate segmentation.

REFERENCES

[1] Jacques Ferlay, Murielle Colombet, Isabelle Soerjomataram, et al., "Cancer statistics for the year 2020: An overview," *International journal of cancer*, vol. 149, no. 4, pp. 778–789, 2021.

[2] Changfa Xia, Xuesi Dong, He Li, Maomao Cao, et al., "Cancer statistics in china and united states, 2022: profiles, trends, and determinants," *Chinese Medical Journal*, vol. 135, no. 05, pp. 584–590, 2022.

[3] Rebecca L. Siegel, Kimberly D. Miller, Hannah E. Fuchs, and Ahmedin Jemal, "Cancer statistics, 2022," *CA: A Cancer Journal for Clinicians*, vol. 72, no. 1, pp. 7–33, 2022.

[4] Jerry A Barbee Jr et al., "Prostate cancer: Facts, causes, and treatments," *Health-System Edition*, vol. 7, no. 2, 2018.

[5] Baris Turkbey, Andrew B. Rosenkrantz, Masoom A. Haider, Anwar R. Padhani, et al., "Prostate imaging reporting and data system version 2.1: 2019 update of prostate imaging reporting and data system version 2," *European Urology*, vol. 76, no. 3, pp. 340–351, 2019.

[6] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 fourth international conference on 3D vision (3DV)*. IEEE, 2016, pp. 565–571.

[7] Geert Litjens, Robert Toth, Wendy van de Ven, Caroline Hoeks, Sjoerd Kerkstra, Bram van Ginneken, et al., "Evaluation of prostate segmentation algorithms for mri: The promise12 challenge," *Medical Image Analysis*, vol. 18, no. 2, pp. 359–373, 2014.

[8] Qikui Zhu, Bo Du, Baris Turkbey, Peter L Choyke, and Pingkun Yan, "Deeply-supervised cnn for prostate segmentation," in *2017 international joint conference on neural networks (IJCNN)*. IEEE, 2017, pp. 178–184.

[9] Geert Litjens, Oscar Debats, Jelle Barentsz, Nico Karssemeijer, and Henkjan Huisman, "Computer-aided detection of prostate cancer in mri," *IEEE transactions on medical imaging*, vol. 33, no. 5, pp. 1083–1092, 2014.

[10] Renato Cuocolo, Albert Comelli, Alessandro Stefano, Viviana Benfante, Navdeep Dahiya, Arnaldo Stanzione, et al., "Deep learning whole-gland and zonal prostate segmentation on a public mri dataset," *Journal of Magnetic Resonance Imaging*, vol. 54, no. 2, pp. 452–459, 2021.

[11] Qinyu Cheni, Bodo Rueckauer, Li Li, Tobi Delbruck, and Shih-Chii Liu, "Reducing latency in a converted spiking video segmentation network," in *2021 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2021, pp. 1–5.

[12] Leonardo Rundo, Changhee Han, Yudai Nagano, Jin Zhang, Ryuichiro Hataya, Carmelo Militello, et al., "Use-net: Incorporating squeeze-and-excitation blocks into u-net for prostate zonal segmentation of multi-institutional mri datasets," *Neurocomputing*, vol. 365, pp. 31–43, 2019.

[13] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," in *Proceedings of CVPR*, 2018, pp. 7132–7141.

[14] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, et al., "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[15] Dimitri Hamzaoui, Sarah Montagne, Raphaele Renard-Penna, Nicholas Ayache, and Hervé Delingette, "Automatic zonal segmentation of the prostate from 2d and 3d t2-weighted mri and evaluation for clinical use," *Journal of Medical Imaging*, vol. 9, no. 2, pp. 024001, 2022.

[16] Haonan Wang, Peng Cao, Jiaqi Wang, and Osmar R Zaiane, "Uctransnet: rethinking the skip connections in u-net from a channel-wise perspective with transformer," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022, vol. 36, pp. 2441–2449.

[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, et al., "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[18] Tsung-Han Tsai and Shih-An Huang, "Live demonstration: Real-time multi-hand segmentation on exhibition," in *2021 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2021, pp. 1–1.

[19] Vinod Nair and Geoffrey E Hinton, "Rectified linear units improve restricted boltzmann machines," in *Icml*, 2010.

[20] Nader Aldoj, Federico Biavati, Florian Michallek, Sebastian Stober, and Marc Dewey, "Automatic prostate and prostate zones segmentation of magnetic resonance images using densenet-like u-net," *Scientific reports*, vol. 10, no. 1, pp. 1–17, 2020.