CAT-Net: A Cross-Slice Attention Transformer Model for Prostate Zonal Segmentation in MRI

Alex Ling Yu Hung, Haoxin Zheng, Qi Miao, Steven S. Raman, Demetri Terzopoulos, *Life Fellow, IEEE*, and Kyunghyun Sung, *Member, IEEE*

***Abstract*— Prostate cancer is the second leading cause of cancer death among men in the United States. The diagnosis of prostate MRI often relies on accurate prostate zonal segmentation. However, state-of-the-art automatic segmentation methods often fail to produce well-contained volumetric segmentation of the prostate zones since certain slices of prostate MRI, such as base and apex slices, are harder to segment than other slices. This difficulty can be overcome by leveraging important multi-scale image-based information from adjacent slices, but current methods do not fully learn and exploit such cross-slice information. In this paper, we propose a novel cross-slice attention mechanism, which we use in a Transformer module to systematically learn cross-slice information at multiple scales. The module can be utilized in any existing deep-learning-based segmentation framework with skip connections. Experiments show that our cross-slice attention is able to capture cross-slice information significant for prostate zonal segmentation in order to improve the performance of current state-of-the-art methods. Cross-slice attention improves segmentation accuracy in the peripheral zones, such that segmentation results are consistent across all the prostate slices (apex, mid-gland, and base). The code for the proposed model is available at https://bit.ly/CAT-Net.**

***Index Terms*— Attention mechanism, deep learning, magnetic resonance imaging, prostate zonal segmentation, transformer network.**

Manuscript received 16 August 2022; accepted 23 September 2022. Date of publication 4 October 2022; date of current version 29 December 2022. This work was supported in part by the National Institutes of Health under Grant R01-CA248506 and in part by the Integrated Diagnostics Program, Departments of Radiological Sciences and Pathology, David Geffen School of Medicine, UCLA. *(Corresponding author: Alex Ling Yu Hung.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the UCLA Institutional Review Board (IRB) with a waiver for written informed consent and is compliant with the 1996 United States Health Insurance Portability and Accountability Act.

Alex Ling Yu Hung and Haoxin Zheng are with the Computer Science Department and the Department of Radiological Sciences, University of California, Los Angeles (UCLA), Los Angeles, CA 90095 USA (e-mail: alexhung96@ucla.edu; haoxinzheng@ucla.edu).

Qi Miao is with the Department of Radiological Sciences, University of California, Los Angeles (UCLA), Los Angeles, CA 90095 USA, and also with the Department of Radiology, The First Affiliated Hospital of China Medical University, Shenyang, Liaoning 110001, China (e-mail: meganmiaoqi@126.com).

Steven S. Raman and Kyunghyun Sung are with the Department of Radiological Sciences, University of California, Los Angeles (UCLA), Los Angeles, CA 90095 USA (e-mail: sraman@mednet.ucla.edu; ksung@mednet.ucla.edu).

Demetri Terzopoulos is with the Computer Science Department, University of California, Los Angeles (UCLA), Los Angeles, CA 90095 USA, and also with VoxelCloud, Inc., Los Angeles, CA 90024 USA (e-mail: dt@cs.ucla.edu).

Digital Object Identifier 10.1109/TMI.2022.3211764

I. Introduction

PROSTATE cancer (PCa) is the most common cancer and the second leading cause of cancer-related death among men in the United States [1]. Multi-parametric MRI (mpMRI), including T2-weighted imaging (T2WI), diffusion-weighted imaging (DWI), and dynamic contrast-enhanced (DCE) MRI, is now the preferred non-invasive imaging technique for prostate cancer (PCa) diagnosis prior to biopsy [2]. According to the Prostate Imaging Reporting and Data System (PI-RADS) [3], the current clinical standard for interpreting mpMRI, a suspicious lesion should be analyzed differently in different prostate zones, among them the transition zone (TZ) and the peripheral zone (PZ), due to variations in image appearance and cancer prevalence [4]. The zonal information is essential and should be provided explicitly for the accurate identification and assessment of suspicious lesions. Moreover, the size of the TZ is often used to evaluate and monitor benign prostate hyperplasia (BPH) in clinical practice [5]. However, the manual annotation of prostate zones is typically time-consuming and highly variable depending on experience level. Therefore, reliable and robust automatic zonal segmentation methods are needed in order to improve PCa detection.

With the emergence of deep learning (DL), DL-based medical image segmentation methods have been proposed to automatically segment the prostate zones [6], [7], [8], [9], [10]. DL-enabled segmentation methods tend to perform well in general, but several studies report that the apex and base locations of the prostate are more difficult to segment than the mid-gland slices [11], [12]. Leveraging information from nearby slices can improve the performance of these methods in all parts of the prostate. However, most 2D-based segmentation methods do not fully consider or systematically learn the available cross-slice information, thus they may disregard important structural information about the prostate, leading to inconsistent segmentation results. It is crucial to take the through-plane information or cross-slice relationship into full consideration when devising prostate zonal segmentation models.

Several 3D DL-based medical image segmentation networks have been proposed in previous studies [13], [14], but their application to prostate MRI has been limited. Following the standard guideline of PI-RADS [3], T2WI images are acquired using the multi-slice 2D Turbo Spin Echo (TSE) sequence, resulting in high in-plane image resolution (e.g., 0.3–1.0 mm) but low through-plane resolution (e.g., 3.0–6.0 mm). Due to the anisotropic nature of the image resolution, existing 3D segmentation networks may not be directly applicable as they are typically designed for nearly isotropic 3D images [13],

[14], and their performance suffers when they are confronted with anistropic data [15], [16], [17].

Transformer networks have become the dominant DL architecture in the natural language processing (NLP) domain as their multi-head self-attention (MHSA) mechanisms learn the global context and the relationship among different word embeddings [18]. Following similar ideas, MHSA mechanisms have recently been applied in the computer vision domain to address tasks such as object detection, segmentation, and classification.

As it can capture the long-range dependencies in images [18], [19], [20], self-attention is a promising mechanism for systematically learning the cross-slice information needed to improve 3D medical image analysis; that is, as compared to convolutional networks whose inherent inductive bias is to focus more on neighboring features. To exploit the fact that prostate zonal segmentation can benefit from information spanning all the slices through the entire prostate instead of only neighboring slices, we devise a 2.5D cross-slice attention-based module that can be incorporated within any network architecture with skip connections in order to capture the global cross-slice relationship. Referring to Fig. 1, we note that a 2.5D method employs only 2D convolutional layers yet learns 3D information, whereas 2D methods also use 2D convolutional layers, but consider only a single image, while 3D methods input 3D volumes and use 3D convolutional layers.

Our main contributions include the following:

1) We formally propose the use of a cross-slice attention mechanism to capture the relationship between MRI slices for the purposes of prostate zonal segmentation.
2) We devise a novel 2.5D Cross-slice Attention Transformer (CAT) module that can be incorporated into existing skip-connection-based network architecture to exploit long-range information from other slices.
3) We perform an experimental study which demonstrates that our proposed DL models, called CAT-Net, are able to improve zonal segmentation results both in general and on different prostate parts, resulting in a new state of the art performance.

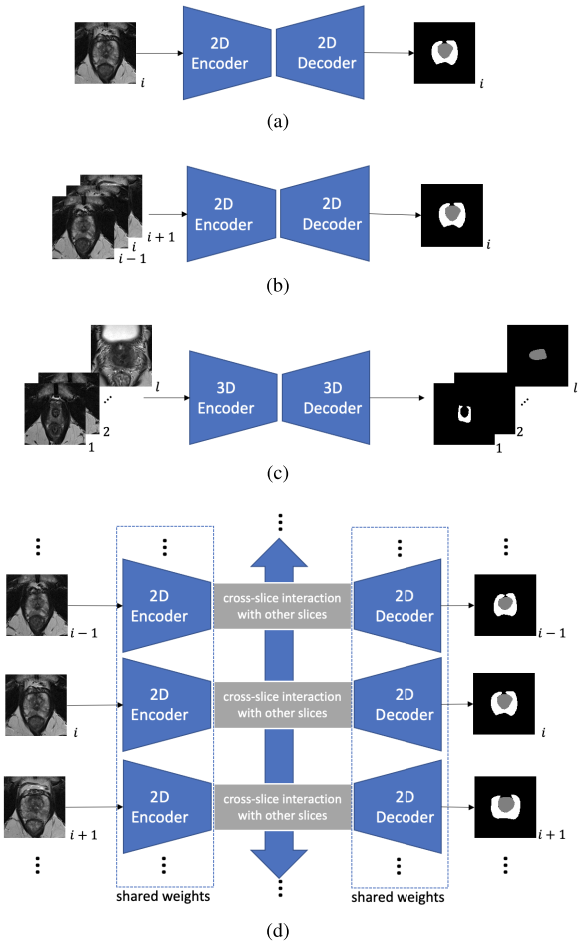

Fig. 1. (a) A conventional 2D segmentation network inputs a slice of interest, performs 2D convolutions, and predicts its segmentation mask. (b) A conventional 2.5D network inputs a middle slice i along with two nearby slices $i-1$ and $i+1$ (defined as 2.5D w/3), performs 2D convolutions, and predicts the segmentation mask of slice i. (c) A conventional 3D network, inputs the entire volume of I slices, performs 3D convolutions, and predicts segmentation masks for all I slices. (d) Our proposed framework employs 2D convolutions, encodes all the slices in the same volume, passes the encoded feature maps from the different slices through the Cross-slice Attention Transformer (CAT) module, and decodes them to predict segmentation maps for all the slices.

II. Related Work

A. Fully Convolutional Network Architectures

The U-Net model [7] has revolutionized medical image segmentation using an encoder-decoder convolutional neural network (CNN) architecture with multi-scale skip connections between the encoder and decoder that preserve high-resolution image information. Building upon U-Net, a number of subsequent segmentation models have been proposed. ResU-Net [21] applied the residual connections from ResNet [22], and the combination of ResNet with U-Net has also proven effective outside of medical image segmentation [23], [24]. Based on ResU-Net, Alom et al. [25] proposed a segmentation network with better feature representation by means of feature accumulation with recurrent convolutional layers. Apart from residual blocks, other attention-based modules have been incorporated into U-Net to improve segmentation; e.g., Rundo et al. [26] incorporated a squeeze and excitation (SE) module [27], a form of channel-wise attention, and Oktay et al. [28] took the attention module [29] a step further and applied it in U-Net. Other U-Net variants have recently emerged. nnU-Net [16] modified the batch normalization [30] in U-Net to instance normalization [31] and employed a leaky rectified linear activation unit (ReLU). Compared to U-Net, U-Net++ [32] used more nested and dense skip connections to better capture the fine-grained details of foreground objects. MSU-Net [33] added multi-scale blocks, which consist of convolutions with different kernel sizes, into U-Net to improve the segmentation details. Gu et al. [34] proposed a novel dense atrous convolution (DAC) block along with a residual multi-kernel pooling (RMP) block and put them in the bottleneck layer of U-Net to capture more high-level features and preserve more spatial information. Apart from U-Net-based network architectures, DRINet [35] consists of dense connection blocks, residual Inception blocks,

and unpooling blocks, which can learn more distinctive features. Crossbar-Net [36] samples vertical and horizonal patches and processes them separately in two sub-models.

Bardis et al. [6] applied the U-Net on T2WI images to perform prostate zonal segmentation. Building on top of DeepLabV3+ [37], Liu et al. [11] employed multi-scale feature pyramid attention (MFPA) to further utilize encoder-side information at different scales to segment prostate zones, and subsequently expanded the network structure with a spatial attentive module (SAM) and Bayesian epistemic uncertainty based on dropout [8]. Cuocolo et al. [38] have thoroughly investigated network structures for prostate zonal segmentation and concluded that ENet [39] is superior to U-Net and ERFNet [40]. Zabihollahy et al. [9] employed two separate networks for the segmentation of different zones and combined them with post-processing. The Dense-2 U-Net [10] was shown to be the best performing 2D deep model for prostate zonal segmentation on the public dataset ProstateX [41]. Not limited to T2WI images, Rundo et al. [42] performed multi-spectral MRI prostate gland segmentation based on clustering.

Three-dimensional CNNs are popular for 3D medical image segmentation. 3D U-Net [13], VNet [43], and DenseVoxel-Net [14] are U-Net architectures that use 3D rather than 2D convolution, and they have proven effective on image data whose cross-pixel distance is similar in all three dimensions. Wang et al. [44] used a two-stage 3D U-Net for multi-modality whole heart segmentation. Other researchers have applied modified 3D U-Nets to infant brain segmentation [45], lung nodule segmentation [46], and brain tumor segmentation [47]. With regard to the application of 3D methods to prostate zonal segmentation, Nai et al. [12] concluded that most methods have similar performance, but that the addition of ADC and DWI data would slightly improve performance. Yu et al. [48] used mixed residual connections in a volumetric ConvNet for whole prostate segmentation. Z-Net [49], capable of capturing more features in a multi-level manner, was built on U-Net to perform 3D prostate segmentation. Wang et al. [50] incorporated group dilated convolution into a deeply supervised fully convolutional framework for prostate segmentation.

B. Transformer Architectures

Originally developed for NLP, the Transformer architecture [18] has recently been gaining momentum in vision. Dosvitskiy et al. [51] proposed the Vision Transformer (ViT) for image classification, treating images as 16×16 words. Unlike the ViT, the Swin Transformer [52] used shifted windows instead of 16×16 fixed-size windows. Carion et al. [53] proposed the Detection Transformer (DETR) for object detection.

Several Transformer-based methods have been proposed for medical image segmentation. MedT [19] coupled a gated position-sensitive axial attention mechanism with a Local-Global (LoGo) training methodology, improving segmentation quality over the U-Net and attention U-Net. UTNet [54] incorporated Transformer blocks into the skip connections in U-Net, allowing the skip connection feature maps to go through a Transformer block before they reach the decoder side. CoTr [55] instead concatenated all the skip connection feature maps and applied the Transformer on the concatenated vector. Petit et al. [56] proposed a U-Net based Transformer framework with a self attention or cross attention module after each skip connection of the normal U-Net. However, these Transformer-based medical image segmentation methods apply attention only between pixels or patches, not between slices.

Our work is closest to that by Guo and Terzopoulos [57], which utilized attention between slices at the bottom layer of a U-Net with a Transformer network. The work demonstrated its feasibility for 3D medical image segmentation but did not consider cross-slice information at different scales and different semantic information learned by multiple heads of the attention mechanism.

III. Methods

A. Overview

We propose a 2.5D Cross-slice Attention Transformer (CAT) module to systematically learn cross-slice information for prostate zonal segmentation. The module is applicable within U-Net-like architectures with skip connections between the encoder and decoder. Segmentation models using CAT modules consist of three parts: a standard 2D encoder, a standard 2D decoder, and CAT modules in different layers. In other words, were we to remove the CAT modules, the network would be a pure 2D network with no interaction between slices. The overall structure of the CAT module is illustrated in Fig. 2, and the incorporation of CAT modules into existing deep models, such as nnU-Net and nnU-Net++, is illustrated in Fig. 3.

The remainder of this section is organized as follows: Section III-B introduces the cross-slice attention mechanism. Cross-slice attention is used in the Transformer block, which is discussed in Section III-C. Positional encoding followed by N Transformer blocks comprise the CAT module, which is described in Section III-D, where we also explain how the CAT module is used in existing networks. We mainly discuss how the CAT module can be incorporated into U-Net and U-Net++ networks to yield our novel CAT-Net models, but any other skip-connection-based networks are also amenable.

B. Cross-Slice Attention

Previous studies have used attention modules to learn inter-channel or inter-pixel relationships, whereas our goal here is to use the attention mechanism to learn the cross-slice relationship for the purposes of our segmentation task. This is important because single-slice prostate zonal segmentation can suffer from ambiguities, especially near apex and base slices. This is also true for manual annotation as clinicians typically refer to nearby slices while annotating the current slice of interest. We devise an algorithm that mimics the manual segmentation process by attending to other slices when the current slice is being annotated. To this end, we regard each slice as being analogous to a word in NLP problems while keeping the spatial information of images intact. After the images are encoded by the encoder, we treat the image features as a deep representation of the "word".

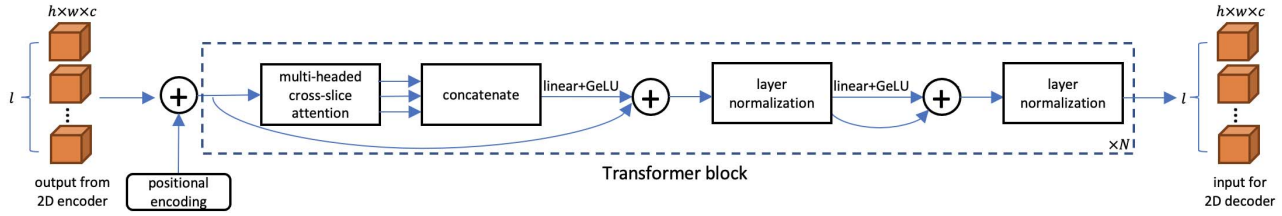

Fig. 2. CAT module includes a positional encoding and *N* Transformer blocks.

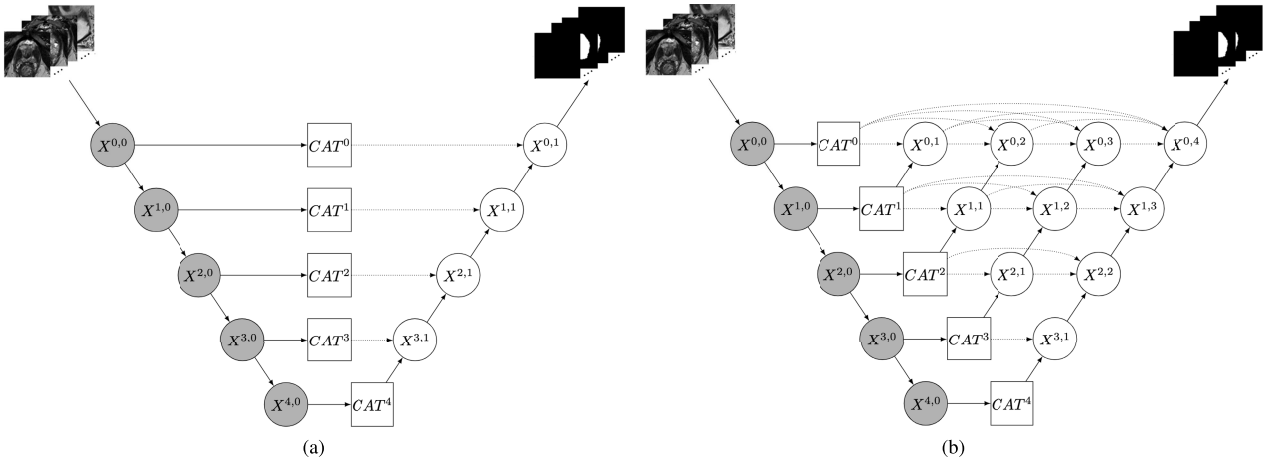

Fig. 3. Implementation of CAT modules in the (a) nnU-Net and (b) nnU-Net++. Following the notation in [32], circles represent feature maps calculated at the corresponding node; gray circles represent the 2D encoder and white circles represent the 2D decoder. Rectangles represent the CAT modules, which operate in 3D. The input stack of images from the patient flow through the encoder, into CAT modules, and then through the decoder to produce the output segmentation.

Let an l-slice stack of input images be represented by the 4D tensor $x \in \mathbb{R}^{l\times h\times w\times c}$, which is a stack of feature maps of height h, width w, and number of channels c. To mitigate computational expense, we leverage the typically high correlation of nearby pixels in feature maps and downsample x by average pooling with a kernel size of k, generating a condensed stack of feature maps $x_{\text{pool}} \in \mathbb{R}^{l\times \frac{h}{k}\times \frac{w}{k}\times c}$. We then calculate queries Q', keys K' as a linear projection of x_{pool}, and values V as a linear projection of x in the channel dimension:

$$Q' = x_{\text{pool}} W_Q, \tag{1}$$
$$K' = x_{\text{pool}} W_K, \tag{2}$$
$$V = x W_V, \tag{3}$$

where $W_Q, W_K, W_V \in \mathbb{R}^{c\times c}$ are learnable weights. Note that Q', K', and V are the deep representation of the l slices and that the above 4D tensor multiplication is defined as follows: The product $C = BW$ of a $l \times h \times w \times c$ tensor B and a $c \times c$ matrix W is calculated as

$$C[i, j, k, m] = \sum_{n=1}^{c} B[i, j, k, n]W[n, m]. \tag{4}$$

The attention matrix $A \in \mathbb{R}^{l\times l}$, which determines how much attention the algorithm pays to other slices while segmenting a slice, is calculated as

$$A = \text{softmax}\left(\frac{QK^T}{\sqrt{hwc/k^2}}\right), \tag{5}$$

where each row of $Q, K \in \mathbb{R}^{l\times \frac{hwc}{k^2}}$, the reshaped matrices of Q' and K', represents the query and key for each slice. The softmax is performed on the second dimension. An element $A[i, j]$ in the attention matrix indicates how similar the query for slice i is to the keys for slice j; i.e., it is computed as a function that performs a weighted average of the values for all the slices to account for the interactions between queries and keys. The output of the cross-slice attention $y \in \mathbb{R}^{l\times h\times w\times c}$ is computed as

$$y = AV. \tag{6}$$

Fig. 4 illustrates our cross-slice attention mechanism, which, unlike the 2D self-attention mechanism [18], uses a condensed deep image feature map to calculate queries Q and keys K and uses a normally encoded feature map for values V, while working in a 4D space to calculate the attention matrix between slices instead of between pixels.

C. Transformer Block

We incorporate our cross-slice attention mechanism into a Transformer block, which is the structure widely used in Transformer-based approaches, where one typically sees a multi-headed attention module followed by linear operations, non-linear modules, and normalizations with skip connections in between. Our Transformer block is shown within the dashed box in Fig. 2. The input to the Transformer block goes through a multi-headed cross-slice attention and is then subjected to linear projections along with non-linear activations. Specifically, we perform multiple cross-slice attention in parallel,

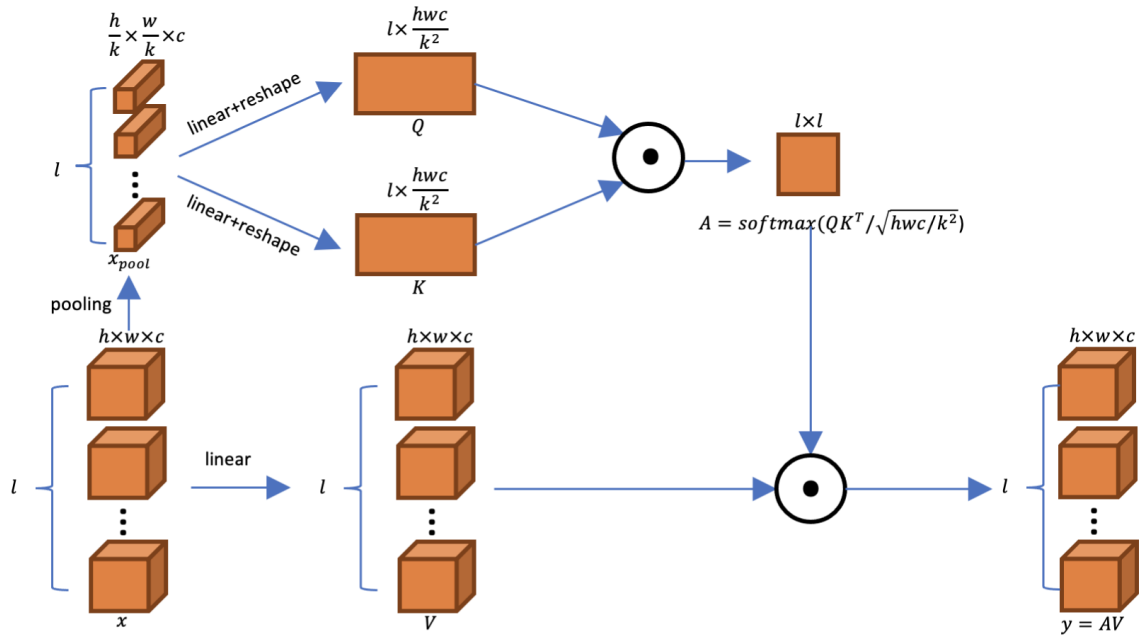

Fig. 4. Cross-slice attention with input x and output y.

obtaining y_i, where $i = 1, 2, \ldots, H$ (H is the number of heads), similar to previous work [18]. This is done so that the network can learn multiple semantics during the attention procedure; i.e., for different meanings, the network learns different attention matrices. Subsequently, we concatenate all the y_i in the c dimension to obtain the output y of the multi-headed cross-slice attention module. The final output of the Transformer block can be expressed as

$$z = \text{Layer\_Norm}(\text{GELU}(z_{\text{int}} W_2 + b_2) + z_{\text{int}}), \tag{7}$$

with intermediate result

$$z_{\text{int}} = \text{Layer\_Norm}(\text{GELU}(y W_1 + b_1) + x), \tag{8}$$

where W_1 and W_2 are the linear projection matrices, b_1 and b_2 are bias terms, GELU is the Gaussian error linear unit [58], and Layer_Norm performs layer normalization [59] across the h, w, and c dimensions.

D. Network Architecture

As shown in Fig. 3, we use nnU-Net [16] and nnU-Net++ [32] as the backbones of our CAT-Net architectures, which we denote as CAT-nnU-Net and CAT-nnU-Net++, respectively. Other network architectures with skip connections between the encoder and decoder would also be suitable. We find that our attention mechanism works well on nnU-Net-like designs where Leaky ReLU and instance normalization is used rather than normal ReLU and batch normalization.

The 2D encoder E takes in a $l \times c_0 \times h_0 \times w_0$ tensor, where l, h_0, w_0, and c_0 are the number of slices, height, width and the number of channels, respectively. It treats l as the batch dimension, where the slices do not interfere with each other. A deep feature map $x_i \in \mathbb{R}^{l\times h_i \times w_i \times c_i}$ is input into the CAT module (Fig. 2), in which it is subjected to a positional encoding and N Transformer blocks. The positional encoding is important in this task, since it tells the network the location of the slice. As in previous work [18], we first add to it a learnable positional encoding PE_i initialized as follows [60]: For all elements on slice p and channel $2j$,

$$PE_i[p, :, :, 2j] = \sin\left(\frac{p}{10000^{2j/c_i}}\right), \tag{9}$$

and for all the elements on slice p and channel $2j+1$,

$$PE_i[p, :, :, 2j+1] = \cos\left(\frac{p}{10000^{2j/c_i}}\right). \tag{10}$$

Then, the feature is passed through N Transformer blocks, where the feature maps of different slices interact to yield the outputs of the skip connection $z_i \in \mathbb{R}^{l\times h_i \times w_i \times c_i}$. Again, we treat the first dimension as the batch dimension, interpreting z_i as l feature maps of 2D images. The 2D decoder takes in z_i and outputs the final segmentation masks.

More generally, encoder E is given l image slices denoted as $x_0 \in \mathbb{R}^{l\times h_0 \times w_0 \times c_0}$ and returns the encoded images $x_i \in \mathbb{R}^{l\times h_i \times w_i \times c_i}$ at different scales $1 \le i \le L$, where L is the total number of layers in the encoder: $\{x_i\}_{1\le i\le L} = E(x_0)$.

We denote the lowest layer in the decoder as D_L and the layers above it as $D_{L-1}, D_{L-2}, \ldots$, and the output of decoder layer i as d_i. In a conventional skip-connection-based network, the decoder takes in the feature maps from different scales:

$$d_i = \begin{cases} D_i(x_i) & \text{if } i = L, \\ D_i(x_i, d_{i+1}) & \text{otherwise,} \end{cases} \tag{11}$$

whereas in our CAT-Net, CAT modules are inserted into the network:

$$d_i = \begin{cases} D_i(\text{CAT}_i(x_i)) & \text{if } i = L, \\ D_i(\text{CAT}_i(x_i), d_{i+1}) & \text{otherwise,} \end{cases} \tag{12}$$

where CAT_i is the CAT module at scale i.

Our approach works better on networks with skip connections at different scales since multi-resolution feature maps capture different semantic information [61], [62]; thus, the attention at different scales should not be the same. The cross-slice attention matrices should be different across the different layers of the network, whereas in networks without skip connections across the different scales, the semantic information at different scales is not represented in the feature maps, and the same attention matrix would be applied, leading to unsatisfactory results.

IV. Experiments

Our experimental study was performed in compliance with the United States Health Insurance Portability and Accountability Act (HIPAA) of 1996 and was approved by the institutional review board (IRB) with a waiver of the requirement for informed consent.

A. Study Population and MRI Data

1) Our Dataset: 296 patients who underwent pre-operative 3 Tesla (3T) mpMRI prior to robotic-assisted laparoscopic prostatectomy were included in the study. Patients with prior radiotherapy or hormonal therapy and with an endorectal coil were excluded. All mpMRI scans were performed on one of the four 3T MRI scanners (Siemens Healthineers MAGNETOM Trio, Skyra, Prisma, and Vida) from January 2013 to December 2018 at a single academic institution. T2WI was acquired using a T2-weighted Turbo Spin Echo (TSE) MR sequence following the standardized imaging protocol of the European Society of Urogenital Radiology (ESUR) PI-RADS guidelines [3]. The T2WI images were used for the zonal segmentation with an in-plane resolution of 0.625 mm^2, a through-plane resolution of 3 mm, and an image size of $320 \times 320 \times 20$ voxels. We cropped the central images to 128×128 and used 238, 29, and 29 patients for training, validation, and testing, respectively. For experiments involving cross validation, we randomized the data and grouped them into five folds.

2) ProstateX Dataset: A total of 193 patients from the ProstateX [41] dataset were included in the study. The imaging was performed by Siemens MAGNETOM Trio and Skyra 3T MR scanners. A turbo spin echo sequence was used to acquire T2WI images, which had a resolution of 0.5 mm in plane and a slice thickness of 3.6 mm. We excluded some data from the original dataset due to differences in image sizes and sequence lengths. The image size is $384 \times 384 \times 18$ voxels, where we pick only the middle 18 slices for each patient. We cropped the central images to 160×160 and resampled them to 128×128 and used 157, 20, and 20 patients for training, validation, and testing, respectively.

B. Implementation Details

To provide input to the models, we normalized the T2WI MRI images; i.e., the normalized intensity of pixel (i, j) on slice k is $\tilde{I}_{i,j,k} = (I_{i,j,k} - \mu)/\sigma$, where $I_{i,j,k}$ is the original pixel intensity and μ and σ are the per-patient mean and standard deviation of the pixel intensity, respectively.

For our nnU-Net implementation, we downsampled five times on the encoder side, and had 64, 128, 256, 512, 1024, and 2048 filters in each of the convolutional layers of the encoder. For our nnU-Net++ implementation, we downsampled four times on the encoder side and had 64, 128, 256, 512, and 1024 filters in each of the convolutional layers of the encoder. The decoders were the exact opposite of the encoders. For the 3D networks, we performed cross-slice upsampling and provided the upsampled volumes as inputs to the 3D models that require more than 2 times downsampling. For a fair comparison against the other U-Net-based models, we ensured that the upsampled volumes could be downsampled at least five times. During testing, only the results from the non-interpolated slices were considered in our comparisons against other models. For other parts of the 3D networks and all the other models in our experiments, we strictly adhered to the architectures described in the original papers.

For data augmentation, we performed only center crop, horizontal flip, and Gamma transform. The same data augmentation scheme was applied across all of the experiments. We applied cross entropy loss as the loss function for 150 epochs in all the training procedures, and used Adam [63] with a learning rate of 0.0001 and weight decay regularization [64] with the parameter set to 1×10^{-5}. We did not perform any post-processing after the segmentation. We maintained a simple training setting in order to elucidate the benefit of our CAT module. In our model, we set the number of Transformer blocks $N = 2$, the number of heads $H = 3$, and the average pooling size $k = 4$. The CAT-Nets were trained on a single Nvidia Quadro RTX 8000 GPU, while the other networks were trained on an Nvidia RTX 3090 GPU.

C. Evaluation

1) Evaluation Metrics: We used Intersection over Union (IoU), Dice coefficient (Dice), Relative absolute volume difference (RAVD), and average symmetric surface distance (ASSD) for evaluation, and these metrics were calculated in a 3D patient-wise manner. For experiments involving statistical testing, we used the Mann-Whitney U Test [65] to statistically test the distribution of results from our model and the competing models.

2) Quantitative Evaluation: We performed a comprehensive comparison between our model and state-of-the-art 2D, 2.5D, and 3D medical image segmentation models.

With regard to 2D methods, we compared ours against DeepLabV3+ [37], the Liu et al. model [8], the Zabihollaghy et al. model [9], CE-Net [34], MSU-Net [33], Dense-2 U-Net [10], nnU-Net++ [32], and nnU-Net [16]. Dense-2 U-Net as well as the models of Liu et al. and Zabihollaghy et al. were designed for zonal segmentation, while the others are generic segmentation models. Specifically, we adopted the 2D nnU-Net from the paper [16]. Additionally, we adapted into U-Net++ the Leaky ReLU along with the instance normalization design, which is the main architectural contribution of nnU-Net, and denote it as nnU-Net++. We did not directly compare against U-Net [7] and U-Net++ because previous work has demonstrated the superiority of nnU-Net-based designs, which is supported by our experimental results in Section IV-C.3.

TABLE I
PERFORMANCE OF NNU-NET BASED MODELS AND OTHER MODELS ON OUR DATASET

		TZ				PZ			
		IoU (%)↑	Dice (%)↑	RAVD (%)↓	ASSD (mm)↓	IoU (%)↑	Dice (%)↑	RAVD (%)↓	ASSD (mm)↓
2D	DeepLabV3+ [37]	78.4	87.8	24.2	0.285	67.4	79.9	38.1	0.479
	Liu et al. [8]	77.7	87.3	25.6	0.292	69.2	81.3	35.9	0.455
	Zabihollahy et al. [9]	80.6	89.1	21.6	0.227	69.2	81.4	38.4	0.640
	CE-Net [34]	78.1	87.5	25.3	0.253	68.5	80.7	36.1	0.524
	MSU-Net [33]	79.8	88.6	22.4	0.214	70.7	82.3	34.1	0.431
	Dense-2 U-Net [10]	80.6	89.2	22.1	0.217	71.1	82.8	32.7	0.252
	nnU-Net w/1 [16]	80.0	88.7	22.4	0.215	71.9	83.4	34.2	0.347
2.5D	nnU-Net w/3	80.7	89.1	22.8	0.206	72.3	83.6	33.5	0.336
	nnU-Net w/5	80.9	89.3	21.9	0.200	72.9	84.1	33.7	0.307
	nnU-Net w/7	79.6	88.5	23.2	0.222	73.4	84.4	30.7	0.291
3D	DenseVoxNet [14]	73.5	84.4	31.1	0.343	61.4	75.4	56.2	0.623
	VNet [43]	78.0	87.4	26.6	0.256	71.0	82.9	35.6	0.302
	3D U-Net [13]	81.5	89.6	20.5	0.188	74.6	85.2	29.7	0.274
	CAT-nnU-Net	**82.6**	**90.4**	**19.3**	**0.175**	**75.8**	**86.1**	**28.2**	**0.259**

TABLE II
PERFORMANCE OF NNU-NET BASED MODELS AND OTHER MODELS ON PROSTATEX

		TZ				PZ			
		IoU (%)↑	Dice (%)↑	RAVD (%)↓	ASSD (mm)↓	IoU (%)↑	Dice (%)↑	RAVD (%)↓	ASSD (mm)↓
2D	DeepLabV3+ [37]	68.9	81.2	38.2	0.734	52.4	68.1	59.0	1.024
	Liu et al. [8]	68.9	81.3	37.9	0.679	52.1	67.9	61.9	1.002
	Zabihollahy et al. [9]	71.0	82.9	34.2	0.609	50.6	66.3	67.4	1.236
	CE-Net [34]	68.0	80.4	38.4	0.849	51.2	67.1	63.9	1.013
	MSU-Net [33]	70.1	82.2	35.0	0.622	53.9	69.2	61.9	0.985
	Dense-2 U-Net [10]	71.1	82.8	36.2	0.590	55.7	70.9	58.5	0.885
	nnU-Net w/1 [16]	70.1	82.1	36.9	0.737	53.5	68.9	58.4	1.055
2.5D	nnU-Net w/3	72.5	83.8	33.2	0.583	55.8	71.0	54.1	0.876
	nnU-Net w/5	71.7	83.1	34.0	0.634	56.2	71.3	54.3	**0.864**
	nnU-Net w/7	71.3	82.9	34.7	0.618	54.3	69.7	57.6	0.970
3D	DenseVoxNet [14]	66.0	78.9	39.7	0.985	52.0	67.6	61.6	1.078
	VNet [43]	71.1	82.9	37.4	0.563	53.9	69.4	57.3	0.961
	3D U-Net [13]	72.6	**83.9**	**32.4**	**0.535**	49.3	64.9	60.6	1.088
	CAT-nnU-Net	**72.7**	**83.9**	**32.4**	0.637	**58.5**	**73.1**	**52.5**	0.890

We continued using nnU-Net and nnU-Net++ as the backbones of our 2.5D models. Like Zhang et al. [66], we stacked nearby slices together and inputted the stacks to 2D networks in order to segment the middle slices. In other words, during each run, the networks take in a stack of images $i-k,\ldots,i,\ldots,i+k$ and output the segmentation mask only for the middle slice i. In our experiments, we stacked 3, 5, and 7 slices together and fed them into the networks, which we denote as nnU-Net and nnU-Net++ w/3, w/5, and w/7. Note that nnU-Net w/1 and nnU-Net++ w/1 are the 2D networks where we only input 1 image per segmentation.

For the 3D methods, we used the most popular segmentation models, DenseVoxNet [14], VNet [43], and 3D U-Net [13].

Table I and Table II compare nnU-Net based architectures on our dataset and ProstateX, respectively, while Table III and Table IV compare nnU-Net++ based architectures on our dataset and ProstateX, respectively. Our model outperforms every other model in almost every metric. Conventional 2.5D methods are better than 2D methods for nnU-Net based models, but performance starts to drop as the method uses a 7-image stack input. For nnU-Net++ based models, 2.5D methods show little or no improvement over 2D methods and the optimal number of slices to include is unclear. Naively stacking nearby slices fails to fully utilize the newly added information from nearby slices without an explicit information exchange mechanism. The best performing 3D methods are usually better than most 2D methods and they are sometimes better than conventional 2.5D methods, but there is no clear-cut winner among the competing methods. Specifically, 3D U-Net is the best performing 3D model on our dataset, but VNet shows better performance on ProstateX. Furthermore, the 3D models generally perform decently in segmenting the PZ in our dataset, but perform poorly on ProstateX. As is shown in the tables, using either nnU-Net or nnU-Net++ as the backbone along with our CAT modules yields better performance than any of the current models. The cross-slice attention in our model enables the network to learn the relationship between slices, which results in better performance.

The preceding tables reveal that, aside from our model, 2.5D methods usually performed the best. However, the optimal number of adjacent slices differed between datasets, prostate zones, and backbone networks. For example, for nnU-Net based models, nnU-Net w/5 performed best on our dataset while nnU-Net w/3 performed best on ProstateX. By contrast,

TABLE III
PERFORMANCE OF NNU-NET++ BASED MODELS AND OTHER MODELS ON OUR DATASET

		TZ				PZ			
		IoU (%)↑	Dice (%)↑	RAVD (%)↓	ASSD (mm)↓	IoU (%)↑	Dice (%)↑	RAVD (%)↓	ASSD (mm)↓
2D	DeepLabV3+ [37]	78.4	87.8	24.2	0.285	67.4	79.9	38.1	0.479
	Liu et al. [8]	77.7	87.3	25.6	0.292	69.2	81.3	35.9	0.455
	Zabihollahy et al. [9]	80.6	89.1	21.6	0.227	69.2	81.4	38.4	0.640
	CE-Net [34]	78.1	87.5	25.3	0.253	68.5	80.7	36.1	0.524
	MSU-Net [33]	79.8	88.6	22.4	0.214	70.7	82.3	34.1	0.431
	Dense-2 U-Net [10]	80.6	89.2	22.1	0.217	71.1	82.8	32.7	0.252
	nnU-Net++ w/1 [32]	82.2	**90.1**	20.5	0.192	75.7	85.9	28.7	0.268
2.5D	nnU-Net++ w/3	81.8	89.8	20.3	0.190	74.8	85.4	30.1	0.265
	nnU-Net++ w/5	81.5	89.7	21.2	0.291	74.9	85.4	29.8	0.287
	nnU-Net++ w/7	81.6	89.8	20.8	0.192	74.5	85.1	28.7	0.269
3D	DenseVoxNet [14]	73.5	84.4	31.1	0.343	61.4	75.4	56.2	0.623
	VNet [43]	78.0	87.4	26.6	0.256	71.0	82.9	35.6	0.302
	3D U-Net [13]	81.5	89.6	20.5	**0.188**	74.6	85.2	29.7	0.274
	CAT-nnU-Net++	**82.3**	**90.1**	**20.0**	**0.188**	**76.1**	**86.3**	**28.3**	**0.255**

TABLE IV
PERFORMANCE OF NNU-NET++ BASED MODELS AND OTHER MODELS ON PROSTATEX

		TZ				PZ			
		IoU (%)↑	Dice (%)↑	RAVD (%)↓	ASSD (mm)↓	IoU (%)↑	Dice (%)↑	RAVD (%)↓	ASSD (mm)↓
2D	DeepLabV3+ [37]	68.9	81.2	38.2	0.734	52.4	68.1	59.0	1.024
	Liu et al. [8]	68.9	81.3	37.9	0.679	52.1	67.9	61.9	1.002
	Zabihollahy et al. [9]	71.0	82.9	34.2	0.609	50.6	66.3	67.4	1.236
	CE-Net [34]	68.0	80.4	38.4	0.849	51.2	67.1	63.9	1.013
	MSU-Net [33]	70.1	82.2	35.0	0.622	53.9	69.2	61.9	0.985
	Dense-2 U-Net [10]	71.1	82.8	36.2	0.590	55.7	70.9	58.5	0.885
	nnU-Net++ w/1 [32]	72.0	83.5	32.9	0.594	56.1	71.1	53.3	0.865
2.5D	nnU-Net++ w/3	72.8	84.0	32.4	0.535	56.5	71.6	55.1	0.847
	nnU-Net++ w/5	71.0	82.7	33.8	0.656	56.7	71.8	53.5	0.855
	nnU-Net++ w/7	71.9	83.3	33.0	0.615	56.8	71.8	57.2	0.874
3D	DenseVoxNet [14]	66.0	78.9	39.7	0.985	52.0	67.6	61.6	1.078
	VNet [43]	71.1	82.9	37.4	0.563	53.9	69.4	57.3	0.961
	3D U-Net [13]	72.6	83.9	32.4	0.535	49.3	64.9	60.6	1.088
	CAT-nnU-Net++	**73.1**	**84.1**	**29.6**	**0.512**	**58.2**	**73.1**	**52.3**	**0.781**

our model consistently performed the best on the different datasets, prostate zones, and backbone networks.

For a more rigorous evaluation, we selected the 2D, 2.5D, and 3D methods in the nnU-Net based comparison that performed best on our dataset (i.e., Table I), which are nnU-Net w/1, nnU-Net w/5, and 3D U-Net, and carried out a 5-fold cross validation to determine the robustness of our method. The results shown in Table V, where the p-values were calculated based on the Mann-Whitney U Test between CAT-nnU-Net and competing models, further establish that our method significantly improves the performance of existing models. Although relative to the 3D U-Net our method does not significantly improve the ASSD in both the TZ and PZ, our CAT module significantly outperforms the 3D U-Net in other metrics. Even though nnU-Net w/5 has better performance than 3D U-Net numerically, the fact that CAT-nnU-Net exhibits more significant improvement over nnU-Net w/5 than 3D U-Net reveals that the performance of 3D U-Net is less consistent.

3) Ablation Study: We conducted ablation studies using our dataset in which we compared the performance of both nn-based and conventional U-Net and U-Net++ architectures with and without CAT modules. Our findings are reported in Table VI and Table VII. CAT modules substantially improved the segmentation of PZ on conventional network architectures. In general, the nn-based networks outperform conventional networks, and incorporating CAT modules into nn-based networks yields the best performance.

In view of the larger improvement in performance when adding CAT modules to nnU-Net and the long training time of CAT-nnU-Net++, we performed an ablation study into the effect of positional encoding and Transformer blocks using only the CAT-nnU-Net. The results on our dataset are reported in Table VIII, where we used 5-fold cross validation and the p-values were calculated based on the Mann-Whitney U Test between the network incorporating all the components and networks missing positional encoding or transformer blocks. We can conclude from the table that using both positional encoding and Transformer blocks can help with prostate zonal segmentation. Using both or either one alone outperforms using neither. There is no statistical significance between using both and using transformer blocks alone in PZ segmentation,

TABLE V
PERFORMANCE OF NNU-NET BASED MODELS AND OTHER MODELS ON OUR DATASET

	TZ				PZ			
	IoU (%)↑	Dice (%)↑	RAVD (%)↓	ASSD (mm)↓	IoU (%)↑	Dice (%)↑	RAVD (%)↓	ASSD (mm)↓
nnU-Net w/1 [16]	79.6***	88.5***	23.6***	0.243***	72.2***	83.6***	32.8***	0.371**
nnU-Net w/5	80.7***	89.2***	22.3***	0.204***	72.7***	84.0***	32.5***	0.354**
3D U-Net [13]	80.3**	89.0**	22.8***	0.209	72.3***	83.6***	32.1**	**0.279***
CAT-nnU-Net	**81.3**	**89.5**	**20.8**	**0.201**	**74.3**	**85.1**	**29.8**	0.301

Label * indicates p-value ≤ 0.1; ** indicates p-value ≤ 0.05; *** indicates p-value ≤ 0.01.

TABLE VI
ABLATION STUDY OF THE EFFECT OF NN-BASED DESIGN AND CAT MODULES ON U-NET

nn	CAT	TZ				PZ			
		IoU (%)↑	Dice (%)↑	RAVD (%)↓	ASSD (mm)↓	IoU (%)↑	Dice (%)↑	RAVD (%)↓	ASSD (mm)↓
		77.5	87.2	25.7	0.311	68.8	80.8	36.1	0.665
✓		80.0	88.7	22.4	0.215	71.9	83.4	34.2	0.347
	✓	75.4	85.4	26.3	0.319	70.5	82.1	33.0	0.397
✓	✓	**82.6**	**90.4**	**19.3**	**0.175**	**75.8**	**86.1**	**28.2**	**0.259**

TABLE VII
ABLATION STUDY OF THE EFFECT OF NN-BASED DESIGN AND CAT MODULES ON U-NET++

nn	CAT	TZ				PZ			
		IoU (%)↑	Dice (%)↑	RAVD (%)↓	ASSD (mm)↓	IoU (%)↑	Dice (%)↑	RAVD (%)↓	ASSD (mm)↓
		77.9	87.4	24.7	0.313	69.0	81.0	35.0	0.507
✓		82.2	**90.1**	20.5	0.192	75.7	85.9	28.7	0.268
	✓	78.8	87.9	23.8	0.256	70.9	82.4	34.4	0.419
✓	✓	**82.3**	**90.1**	**20.0**	**0.188**	**76.1**	**86.3**	**28.3**	**0.255**

TABLE VIII
ABLATION STUDY OF THE EFFECT OF POSITIONAL ENCODING AND TRANSFORMER BLOCKS

Positional Encoding	Transformer Blocks	TZ				PZ			
		IoU (%)↑	Dice (%)↑	RAVD (%)↓	ASSD (mm)↓	IoU (%)↑	Dice (%)↑	RAVD (%)↓	ASSD (mm)↓
		79.6***	88.5**	23.6***	0.243***	72.2***	83.6***	32.8***	0.371**
✓		80.5**	89.0*	23.0**	0.220*	73.1***	84.2***	32.3***	0.350**
	✓	80.6*	89.1*	21.8**	0.202***	**74.4**	**85.1**	30.7	0.302
✓	✓	**81.3**	**89.5**	**20.8**	**0.201**	74.3	**85.1**	**29.8**	**0.301**

Label * indicates p-value ≤ 0.1; ** indicates p-value ≤ 0.05; *** indicates p-value ≤ 0.01.

but using both significantly outperforms using only Transformer blocks in TZ segmentation. Using both is also better than using just positional encoding.

4) Qualitative Evaluation: Some representative results on our dataset are shown in Fig. 5. In Fig. 5a, we compare our CAT-nnU-Net with other nnU-Net-based 2D and 2.5D models. The other nnU-Net-based models tend to overpredict PZ on the side of the TZ.

In the example of Fig. 5a, our model performs better in the segmentation of PZ, especially in the lower middle region and the two upper head regions of the PZ. Without effectively considering the cross-slice relationship, e.g., where the PZ starts to emerge, the models would poorly segment the upper head region of the PZ. This phenomenon is clearly revealed by the results of the other models, which over-predict the PZ region with many false positive predictions. Additionally, when including 7 slices in the 2.5D segmentation, the segmentation of the lower part of the PZ produces spurious spur-like regions. As the arrows indicate in Fig. 5a, there are obvious errors in the segmentation, but the segmentation produced by our method does not suffer from such errors.

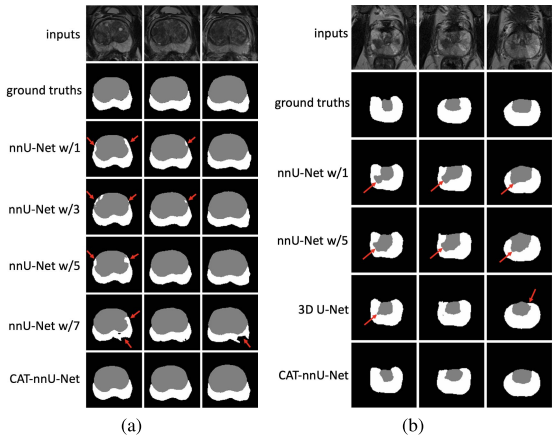

Fig. 5. Comparison of the CAT-nnU-Net against (a) other nnU-Net based models and (b) nnU-Net w/1, nnU-Net w/5, and 3D U-Net. The segmentation masks of PZ are shown in white, those of TZ in gray, and those of other tissues in black.

TABLE IX
PERFORMANCE WHEN USING CAT MODULES ONLY IN LAYERS 0–3 AND IN ALL 6 LAYERS

	TZ				PZ			
	IoU (%)↑	Dice (%)↑	RAVD (%)↓	ASSD (mm)↓	IoU (%)↑	Dice (%)↑	RAVD (%)↓	ASSD (mm)↓
CAT in Layers 0–3	82.4	90.2	19.7	0.188	**76.3**	**86.4**	**27.4**	**0.244**
CAT in all 6 layers	**82.6**	**90.4**	**19.3**	**0.175**	75.8	86.1	28.2	0.259

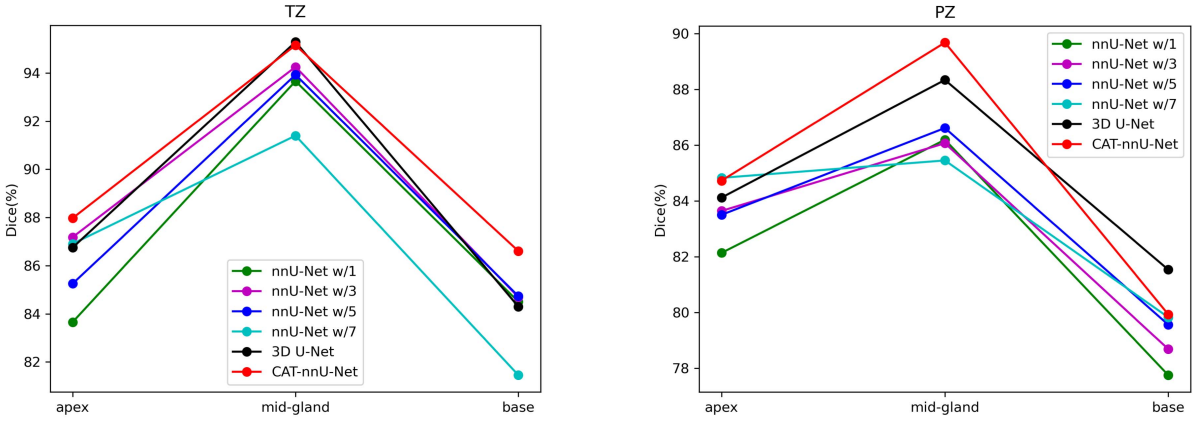

Fig. 6. Performance of (a) TZ and (b) PZ segmentation on different prostate parts by different algorithms.

For a clearer demonstration of CAT-nnU-Net's superiority, in Fig. 5b we compare it against the best performing 2D and 2.5D nnU-Net based methods and the best 3D method based on the quantitative results, which are nnU-Net w/1, nnU-Net w/5, and 3D U-Net. As is indicated by the arrows, other models tend to overpredict the TZ segmentation. Moreover, the 3D U-Net also has some problems in segmenting the upper head region of the PZ. By contrast, our method can perform accurate segmentation of the TZ without predicting unrealistic shape for it, and around the upper head region of the PZ.

5) Evaluation on Different Parts of the Prostate: Prostate zonal segmentation approaches exhibit significant performance differences in different parts of the prostate [8]; hence, we will next investigate the performance of our method on different parts of the prostate on our dataset. Defining the first and last three slices of the prostate as apex and base, respectively, and the remaining slices are mid-gland slices, we continue our evaluation against other nnU-Net based models and 3D U-Net.

As shown in Fig. 6, other models perform differently in the different prostate zones. Among the other nnU-Net-based models, nnU-Net w/5 is the best for TZ segmentation in all parts based on the quantitative results, while it performs badly in apex slices. Although nnU-Net w/7 performs well in segmenting PZ in the apex and base slices, its performance in the mid-gland slices and TZ segmentation is underwhelming. Therefore, it would be hard to determine how many slices to include when performing traditional 2.5D prostate zonal segmentation. However, our method inputs all the slices and learns the relationship between slices. 3D U-Net performs well in PZ segmentation on base slices and is on par with our method in TZ segmentation on mid-gland slices, but the performance of 3D U-Net diminishes elsewhere. Although our method yields a marginally worse result than the nnU-Net w/7 on apex in the PZ and the 3D U-Net on base in the PZ, it clearly outperforms the other methods in general. Our method is consistently among the top two best performing methods in every scenario, whereas the performances of the other methods decrease in certain scenarios. Note that performance comparisons across different parts of the prostate may have practical limitations because of ground-truth annotation complexities due to the inherent ambiguity of zonal appearance at the apex and base slices, which manifests as high inter-reader variability [8].

Fig. 7 shows qualitative results of prostate zonal segmentation on a single patient. Among the models, our CAT-nnU-Net is the most consistent, whereas other 2D and 2.5D models suffer from over-prediction and inconsistency in anatomical structures.

6) Understanding the Attention Matrices: Since CAT-nnU-Net yields a larger improvement over nnU-Net than CAT-nnU-Net++ yields over nnU-Net++, we used the former in the following experiment. The most notable attention matrices A from each layer are visualized in Fig. 8, ranging from fine-resolution, top Layer 0 to coarse-resolution, bottom Layer 5. The CAT modules in the finer Layers 0–3 were able to learn something meaningful, while those in the coarser Layers 4 and 5 learned nothing. This observation is supported by the segmentation performance results reported in Table IX, which are similar with and without attention in Layers 4 and 5. The full network yields slightly better results on TZ but it has worse results on PZ. The cross-slice attention in the coarsest layers may confuse the network when segmenting the more challenging PZ. Apparently, the cross-slice attention mechanism is unable to learn anything useful in Layers 4 and 5 because the network need not rely on nearby slices to learn coarse information. This seems to be consistent with how clinicians annotate prostate images, as they need to refer to nearby slices only to segment the finer details.

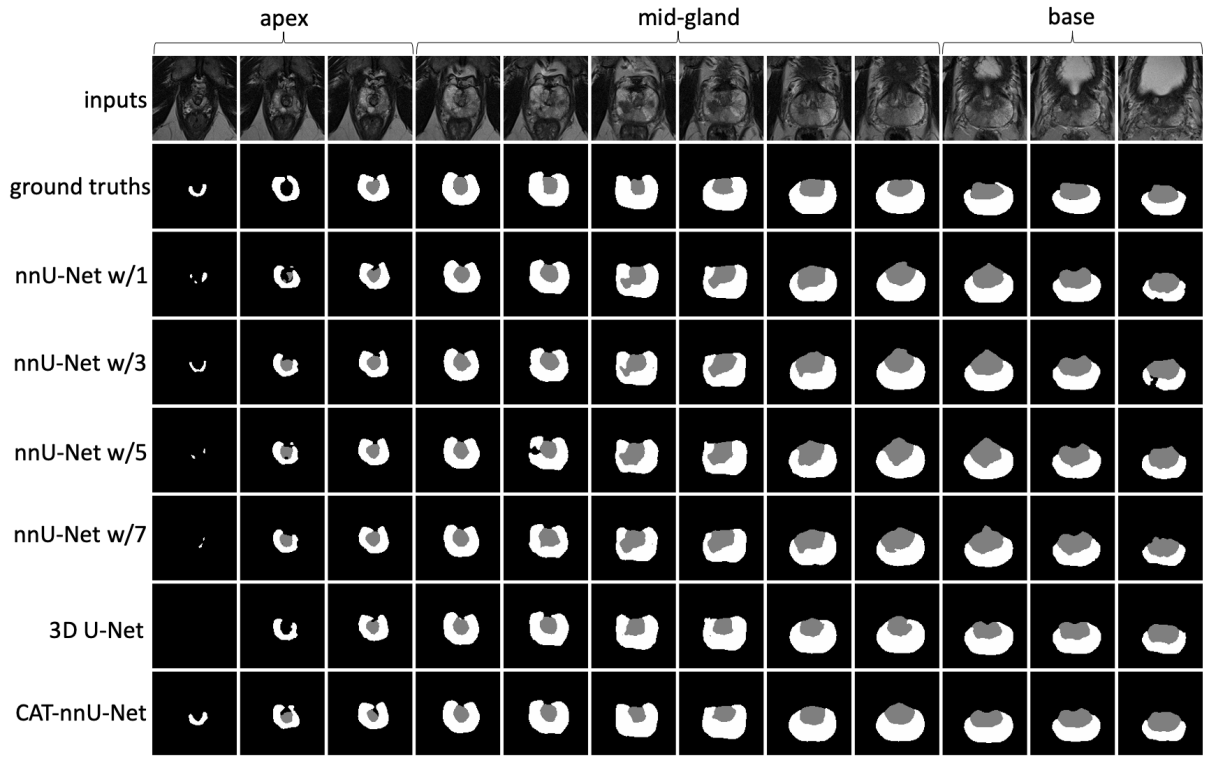

Fig. 7. Comparison of our CAT-nnU-Net against other U-Net based models on different prostate parts.

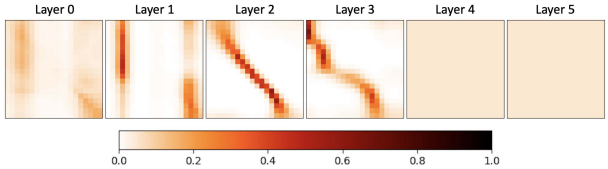

Fig. 8. Attention matrices A from Layers 0–5. Each element $A[i,j]$ indicates the attention slice i pays to slice j, with darker pixels denoting higher values, and $\sum_j A[i,j] = 1$.

Though the CAT modules in Layers 4 and 5 learn nothing useful, we include them nonetheless, since doing so does not significantly hurt performance. The attention matrices in Layers 0 and 1 indicate that for the details, mid-gland slices will attend more to either the apex or the base slices, and slices in the apex and base will attend more to the slices in their own prostate part. At the same time, apex and base slices will also attend slightly to each other, which shows that there are some long-range dependencies in the segmentation of finer details. The attention in Layers 2 and 3, which contain more significant information than Layers 0 and 1, focuses more on nearby slices in the mid-gland slices. This is in line with the results of the previous section, where more slices are needed to accurately segment the base and apex, hence the more blurry regions at the top left and bottom right of Layer 2 as well as the bottom right of Layer 3.

V. Discussion

Our study demonstrated that applying the CAT modules on skip connections at different scales can improve the performance of prostate zonal segmentation by systematically exploiting cross-slice attention. In particular, we see a significant improvement in PZ segmentation compared with TZ segmentation. We believe that this may be because it is relatively easier to segment the TZ without accounting for nearby slices as the shape of the TZ is well-defined and consistent across subjects. By contrast, PZ segmentation is harder than TZ segmentation since the shape of the PZ is less clear in certain slices, which may be why our cross-slice attention module improved the performance of PZ segmentation relative to existing networks.

Our cross-slice attention module has shown its superiority to other methods for the following reasons: 2D networks cannot acquire useful information from nearby slices. 3D convolution approaches need to work with interpolated data due to the anisotropy of the MRI data (i.e., through-plane resolution substantially lower than in-plane resolution). This can be problematic if the number of slices is large. The problem with conventional 2.5D methods may be that the network cannot sufficiently process which slice is the one to be segmented. Additionally, including more slices in the stack introduces more useless information, so the network could have a harder time learning the useful information. Furthermore, inputting more slices might make the network more susceptible to overfitting.

We find that CAT modules yield the biggest improvement on skip-connection-based encoder-decoder networks, e.g., U-Net and U-Net++ using leaky ReLU as activation and instance normalization, which is the scheme in nnU-Net. However, even with different activation functions and normalization schemes, performance drops when applying the cross-slice attention on frameworks without skip connections, e.g., DeepLabV3+ and the model of Liu et al. The networks can learn different attention at different resolutions since the information in the feature maps at different scales differs. This enables the

network to learn more meaningful cross-slice information. On the other hand, the network could be confused by the cross-slice information without skip connections at different scales, leading to worse performance. Further investigation into why the CAT modules do not work on architectures lacking skip connections and how to make similar ideas work on these architectures would be a good direction for further study.

We only considered T2WI, but in practice performing prostate segmentation on different MRI or multispectral MRI can potentially improve segmentation performance [12], [42]. Other image contrasts, such as T1WI, are capable of providing information related to the segmentation that is not present in T2WI [67]. To perform multispectral MRI tasks within our framework, we can add the multispectral MRI images as different input channels to the network, where each channel is one type of MRI image. However, further consideration may be needed for zonal segmentation since T1WI generally does not contain good tissue contrast between prostate zones.

We evaluated our approach on two datasets: our dataset and the ProstateX dataset. In absolute terms, its performance differed between the two datasets, but our CAT module consistently improved performance on both datasets in relative terms. Although our method was evaluated only on prostate zonal segmentation due to the limited data available, it promises to be applicable in other segmentation tasks where information in nearby slices is needed. This is most useful in anisotropic imaging data where cross-slice information is crucial to accurate segmentation since 3D CNNs either tend to underperform or must work with interpolated data having high memory requirements. For isotropic data, our method would still produce decent results compared to other 2D and 2.5D methods, albeit with high memory requirements since our method is less memory efficient than those lacking CAT modules. For instance, for inputs of size $128 \times 128 \times 20$, CAT-nnU-Net has 6.1×10^8 trainable parameters compared with 1.4×10^8 for nnU-Net, and CAT-nnU-Net++ has 3.9×10^8 trainable parameters as opposed to 3.7×10^7 for U-Net++. How to adapt our method to problems other than prostate zonal segmentation and how to make it more memory efficient would be interesting directions for future research.

VI. Conclusion

We have demonstrated improved prostate zonal segmentation by applying a self-attention mechanism to systematically learn and leverage cross-slice information. More specifically, we have proposed a Cross-slice Attention Transformer (CAT) module and incorporated it into state-of-the-art 2D skip-connection-based deep networks. Our resulting CAT-Net models perform better than the current state-of-the-art 2D, 2.5D, and 3D competitors in the task of prostate zonal segmentation, especially in the peripheral zone (PZ) of the prostate. Compared with conventional 2.5D prostate segmentation methods, our segmentation performance was good across the apex, mid-gland, and base slices. Furthermore, our analysis of the attention matrices provided insights into how the cross-slice attention mechanism helps in prostate segmentation. Our approach has proven to be useful in skip-connection-based networks like U-Net and U-Net++, and our ablation study has shown that each of its components contributes such that their combination yields the best results. Further research is needed to investigate the optimal incorporation of our CAT modules into other network architectures.

References

[1] P. Rawla, "Epidemiology of prostate cancer," *World J. Oncol.*, vol. 10, no. 2, p. 63, 2019.

[2] M. B. Appayya et al., "National implementation of multi-parametric magnetic resonance imaging for prostate cancer detection—Recommendations from a U.K. Consensus meeting," *BJU Int.*, vol. 122, no. 1, p. 13, 2018.

[3] B. Turkbey et al., "Prostate imaging reporting and data system version 2.1: 2019 update of prostate imaging reporting and data system version 2," *Eur. Urol.*, vol. 76, pp. 340–351, Sep. 2019.

[4] B. Israel, M. V. D. Leest, M. Sedelaar, A. R. Padhani, P. Zámecnik, and J. O. Barentsz, "Multiparametric magnetic resonance imaging for the detection of clinically significant prostate cancer: What urologists need to Know. Part 2: Interpretation," *Eur. Urol.*, vol. 77, no. 4, pp. 469–480, Apr. 2020.

[5] N. Lawrentschuk, G. Ptasznik, and S. Ong, "Benign prostate disorders," in *Endotext [Internet]*, K. R. Feingold et al., Eds. South Dartmouth, MA, USA: MDText.com, Oct. 2021. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK279008/

[6] M. Bardis et al., "Segmentation of the prostate transition zone and peripheral zone on MR images with deep learning," *Radiol., Imag. Cancer*, vol. 3, no. 3, May 2021, Art. no. e200024.

[7] 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.*, Munich, Germany, in Lecture Notes in Computer Science, vol. 9349. Cham, Switzerland: Springer, 2015, pp. 234–241.

[8] Y. Liu et al., "Exploring uncertainty measures in Bayesian deep attentive neural networks for prostate zonal segmentation," *IEEE Access*, vol. 8, pp. 151817–151828, 2020.

[9] F. Zabihollahy, N. Schieda, S. K. Jeyaraj, and E. Ukwatta, "Automated segmentation of prostate zonal anatomy on T2-weighted (T2W) and apparent diffusion coefficient (ADC) map MR images using U-Nets," *Med. Phys.*, vol. 46, pp. 3078–3090, Jul. 2019.

[10] N. Aldoj, F. Biavati, F. Michallek, S. Stober, and M. Dewey, "Automatic prostate and prostate zones segmentation of magnetic resonance images using DenseNet-like U-Net," *Sci. Rep.*, vol. 10, no. 1, pp. 1–17, Dec. 2020.

[11] Y. Liu et al., "Automatic prostate zonal segmentation using fully convolutional network with feature pyramid attention," *IEEE Access*, vol. 7, pp. 163626–163632, 2019.

[12] Y.-H. Nai et al., "Evaluation of multimodal algorithms for the segmentation of multiparametric MRI prostate images," *Comput. Math. Methods Med.*, vol. 2020, pp. 1–12, Oct. 2020.

[13] O. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning dense volumetric segmentation from sparse annotation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, Athens, Greece, in Lecture Notes in Computer Science, vol. 9902. Cham, Switzerland: Springer, 2016, pp. 424–432.

[14] L. Yu et al., "Automatic 3D cardiovascular MR segmentation with densely-connected volumetric convnets," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, Quebec City, QC, Canada, in Lecture Notes in Computer Science, vol. 10435. Cham, Switzerland: Springer, 2017, pp. 287–295.

[15] H. Jia et al., "3D APA-Net: 3D adversarial pyramid anisotropic convolutional network for prostate segmentation in MR images," *IEEE Trans. Med. Imag.*, vol. 39, no. 2, pp. 447–457, Feb. 2020.

[16] F. Isensee et al., "NnU-Net: Self-adapting framework for U-Net-based medical image segmentation," 2018, *arXiv:1809.10486*.

[17] F. Isensee, P. F. Jaeger, P. M. Full, I. Wolf, S. Engelhardt, and K. H. Maier-Hein, "Automatic cardiac disease assessment on cine-MRI via time-series segmentation and domain specific features," in *Proc. Int. Workshop Stat. Atlases Comput. Models Heart*, 2017, pp. 120–129.

[18] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[19] J. Maria Jose Valanarasu, P. Oza, I. Hacihaliloglu, and V. M. Patel, "Medical transformer: Gated axial-attention for medical image segmentation," 2021, *arXiv:2102.10662*.

[20] G. Tang, M. Müller, A. Rios, and R. Sennrich, "Why self-attention? A targeted evaluation of neural machine translation architectures," 2018, *arXiv:1808.08946*.

[21] A. Khanna, N. D. Londhe, S. Gupta, and A. Semwal, "A deep residual U-Net convolutional neural network for automated lung segmentation in computed tomography images," *Biocybern. Biomed. Eng.*, vol. 40, no. 3, pp. 1314–1327, Jul. 2020.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[23] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual U-Net," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 5, pp. 749–753, May 2018.

[24] F. I. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu, "ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data," *ISPRS J. Photogramm. Remote Sens.*, vol. 162, pp. 94–114, Apr. 2020.

[25] M. Z. Alom, C. Yakopcic, M. Hasan, T. M. Taha, and V. K. Asari, "Recurrent residual U-Net for medical image segmentation," *J. Med. Imag.*, vol. 6, no. 1, 2019, Art. no. 014006.

[26] L. Rundo 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, Nov. 2019.

[27] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[28] O. Oktay et al., "Attention U-Net: Learning where to look for the pancreas," 2018, *arXiv:1804.03999*.

[29] S. Jetley, N. A. Lord, N. Lee, and P. H. S. Torr, "Learn to pay attention," 2018, *arXiv:1804.02391*.

[30] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. 32nd Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[31] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," 2016, *arXiv:1607.08022*.

[32] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-Net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support* (Lecture Notes in Computer Science), vol. 11045. Cham, Switzerland: Springer, 2018, pp. 3–11.

[33] R. Su, D. Zhang, J. Liu, and C. Cheng, "MSU-Net: Multi-scale U-net for 2D medical image segmentation," *Frontiers Genet.*, vol. 12, p. 140, Feb. 2021.

[34] Z. Gu et al., "Ce-Net: Context encoder network for 2D medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 38, no. 10, pp. 2281–2292, Oct. 2019.

[35] L. Chen, P. Bentley, K. Mori, K. Misawa, M. Fujiwara, and D. Rueckert, "DRINet for medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 37, no. 11, pp. 2453–2462, Nov. 2018.

[36] Q. Yu, Y. Shi, J. Sun, Y. Gao, J. Zhu, and Y. Dai, "Crossbar-Net: A novel convolutional neural network for kidney tumor segmentation in CT images," *IEEE Trans. Image Process.*, vol. 28, no. 8, pp. 4060–4074, Aug. 2019.

[37] 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.

[38] R. Cuocolo et al., "Deep learning whole-gland and zonal prostate segmentation on a public MRI dataset," *J. Magn. Reson. Imag.*, vol. 54, no. 2, pp. 452–459, 2021.

[39] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, "ENet: A deep neural network architecture for real-time semantic segmentation," 2016, *arXiv:1606.02147*.

[40] E. Romera, J. M. Álvarez, L. M. Bergasa, and R. Arroyo, "ERFNet: Efficient residual factorized ConvNet for real-time semantic segmentation," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 1, pp. 263–272, Jan. 2018.

[41] 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, Jan. 2014.

[42] L. Rundo et al., "Fully automatic multispectral MR image segmentation of prostate gland based on the fuzzy c-means clustering algorithm," in *Multidisciplinary Approaches to Neural Computing* (Smart Innovation, Systems and Technologies), vol. 69. Cham, Switzerland: Springer, 2018, pp. 23–37.

[43] 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)*, Oct. 2016, pp. 565–571.

[44] C. Wang, T. MacGillivray, G. Macnaught, G. Yang, and D. Newby, "A two-stage 3D UNet framework for multi-class segmentation on full resolution image," 2018, *arXiv:1804.04341*.

[45] S. Qamar, H. Jin, R. Zheng, P. Ahmad, and M. Usama, "A variant form of 3D-UNet for infant brain segmentation," *Future Gener. Comput. Syst.*, vol. 108, pp. 613–623, Jul. 2020.

[46] Z. Xiao, B. Liu, L. Geng, F. Zhang, and Y. Liu, "Segmentation of lung nodules using improved 3D-UNet neural network," *Symmetry*, vol. 12, no. 11, p. 1787, Oct. 2020.

[47] J. Chang et al., "Brain tumor segmentation based on 3D Unet with multi-class focal loss," in *Proc. 11th Int. Congr. Image Signal Process., Biomed. Eng. Informat. (CISP-BMEI)*, Oct. 2018, pp. 1–5.

[48] 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. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 66–72.

[49] Y. Zhang, J. Wu, W. Chen, Y. Chen, and X. Tang, "Prostate segmentation using Z-Net," in *Proc. IEEE 16th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2019, pp. 11–14.

[50] B. Wang et al., "Deeply supervised 3D fully convolutional networks with group dilated convolution for automatic MRI prostate segmentation," *Med. Phys.*, vol. 46, no. 4, pp. 1707–1718, Apr. 2019.

[51] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[52] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," 2021, *arXiv:2103.14030*.

[53] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.*, Glasgow, U.K., in Lecture Notes in Computer Science, vol. 12346. Cham, Switzerland: Springer, 2020, pp. 213–229.

[54] Y. Gao, M. Zhou, and D. N. Metaxas, "UTNet: A hybrid transformer architecture for medical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, Strasbourg, France, in Lecture Notes in Computer Science, vol. 12908. Cham, Switzerland: Springer, 2021, pp. 61–71.

[55] Y. Xie, J. Zhang, C. Shen, and Y. Xia, "CoTr: Efficiently bridging CNN and transformer for 3D medical image segmentation," 2021, *arXiv:2103.03024*.

[56] O. Petit, N. Thome, C. Rambour, L. Themyr, T. Collins, and L. Soler, "U-Net transformer: Self and cross attention for medical image segmentation," in *Proc. Int. Workshop Mach. Learn. Med. Imag.*, in Lecture Notes in Computer Science, Strasbourg, France, vol. 12966. Cham, Switzerland: Springer, 2021, pp. 267–276.

[57] D. Guo and D. Terzopoulos, "A transformer-based network for anisotropic 3D medical image segmentation," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 8857–8861.

[58] D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUs)," 2016, *arXiv:1606.08415*.

[59] J. Lei Ba, J. Ryan Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*.

[60] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin, "Convolutional sequence to sequence learning," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 1243–1252.

[61] L. A. Gatys, A. S. Ecker, and M. Bethge, "A neural algorithm of artistic style," 2015, *arXiv:1508.06576*.

[62] C. Olah, A. Mordvintsev, and L. Schubert, "Feature visualization," *Distill*, 2017. [Online]. Available: https://distill.pub/2017/feature-visualization, doi: 10.23915/distill.00007.

[63] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[64] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2017, *arXiv:1711.05101*.

[65] N. Nachar, "The Mann-Whitney U: A test for assessing whether two independent samples come from the same distribution," *Tuts. Quant. Methods Psychol.*, vol. 4, no. 1, pp. 13–20, Mar. 2008.

[66] H. Zhang et al., "Multiple sclerosis lesion segmentation with Tiramisu and 2.5D stacked slices," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, Shenzhen, China, in Lecture Notes in Computer Science, vol. 11764. Cham, Switzerland: Springer, 2019, pp. 338–346.

[67] S. Ozer et al., "Supervised and unsupervised methods for prostate cancer segmentation with multispectral MRI," *Med. Phys.*, vol. 37, no. 4, pp. 1873–1883, 2010.