CSAM: A 2.5D Cross-Slice Attention Module for Anisotropic Volumetric Medical Image Segmentation

Alex Ling Yu Hung Haoxin Zheng Kai Zhao Xiaoxi Du Kaifeng Pang Qi Miao
Steven S. Raman Demetri Terzopoulos Kyunghyun Sung
University of California, Los Angeles

Abstract

A large portion of volumetric medical data, especially magnetic resonance imaging (MRI) data, is anisotropic, as the through-plane resolution is typically much lower than the in-plane resolution. Both 3D and purely 2D deep learning-based segmentation methods are deficient in dealing with such volumetric data since the performance of 3D methods suffers when confronting anisotropic data, and 2D methods disregard crucial volumetric information. Insufficient work has been done on 2.5D methods, in which 2D convolution is mainly used in concert with volumetric information. These models focus on learning the relationship across slices, but typically have many parameters to train. We offer a Cross-Slice Attention Module (CSAM) with minimal trainable parameters, which captures information across all the slices in the volume by applying semantic, positional, and slice attention on deep feature maps at different scales. Our extensive experiments using different network architectures and tasks demonstrate the usefulness and generalizability of CSAM. Associated code is available at https://github.com/aL3x-O-o-Hung/CSAM.

1. Introduction

Deep learning (DL) based algorithms have advanced computational medical image analysis by virtue of their superior performance relative to conventional approaches [1, 25]. In particular, convolutional neural networks (CNNs), such as U-Net [34], U-Net++ [51], and MSU-Net [36], have achieved state-of-the-art 2D medical image segmentation results.

Radiologists typically analyze tomographic images as volumes, attending to multiple image slices to make accurate medical decisions. 2D image segmentation methods, including the aforecited ones, disregard information across slices. Methods using 3D convolutions [7, 27] perform volumetric analysis, but their performance is limited when confronted by anisotropic image data [15, 17, 18, 20]. For example, in T2-weighted MRI the through-plane resolution (3–6 mm) is typically one-third to one-tenth of the in-plane resolution (0.3–1.0 mm) [26]. Furthermore, the gaps between slices and the through-plane resolution can differ for different organs, so one must either upsample the volumes or customize the deep networks to specific datasets and applications.

Other methods have been proposed that use only slice-based 2D convolution, but incorporate volumetric information via the 2D feature maps, instead of directly using 3D convolution on the volume. Zhang et al. [49] and Han et al. [11] stacked nearby slices together as the input to networks, while Yu et al. [47] and Wang et al. [42] combined the input images in the volume in more sophisticated ways. Attention-based methods, such as RsaNet [48], SAU-Net [50], AFTer-U-Net [46], SATr [23], that of Guo and Terzopoulos [10] (hereafter denoted GT-Net), and our CAT-Net [15], use cross-slice attention mechanisms to learn the relationship between different slices. Although these methods are promising, it is hard to know the optimal number of slices to use in the input stack or include in the attention, and it is difficult to train a large number of parameters in the Transformer blocks. Despite the challenges, this category of methods, loosely called 2.5D methods, appears to be a better at dealing with anisotropic volumetric medical images. We will provide a formal definition of 2.5D medical image segmentation elsewhere in this paper.

Instead of using a Transformer network, we propose a Cross-Slice Attention Module (CSAM) for 2.5D medical image segmentation, which significantly reduces the number of trainable parameters relative to other 2.5D methods. Unlike 2.5D segmentation methods that necessitate determining the optimal number of neighboring slices to include, CSAM captures the global cross-slice information across all the slices through its attention mechanism on multi-scale deep feature maps that incorporates the cross-slice semantic and positional attention as well as the importance of the feature maps from different slices. Additionally, we model the uncertainty across slices to better regularize the segmentation.

The main contributions of this paper are (1) a formal definition of 2.5D medical image segmentation models, (2) a new 2.5D cross-slice attention mechanism that greatly re-

duces the number of parameters compared to the current state-of-the-art 2.5D models, (3) the CSAM that is conveniently insertable into existing 2D CNN-based networks to enable volumetric image segmentation, (4) incorporation of our CSAMs into different backbone networks, and (5) extensive validation studies on prostate, placenta, and cardiac MRI segmentation, demonstrating improved model performance over corresponding 2D, 3D, and the previous state-of-the-art 2.5D methods.

2. Related Work

The U-Net [34] encoder-decoder architecture with skip connections, which preserve detailed, high-resolution, and semantic information, has revolutionized medical image segmentation as countless published models are based on or extend this architecture. A fruitful research avenue focuses on combining U-Net with other modules. ResU-Net [21] introduced the residual connections from ResNet [13], Rundo et al. [35] incorporated the squeeze and excitation (SE) module [14], and Oktay et al. [29] injected the widely used attention mechanism [39] into U-Net. Other U-Net variants have also emerged. nnU-Net [17] replaced the batch normalization [16] and rectified linear activation unit (ReLU) with instance normalization [37] and leaky ReLU. To better capture the fine-grained details of the object of interest, U-Net++ uses more nested and dense skip connections between the encoder and decoder. MSU-Net [36] added convolutions with different kernel sizes to capture multi-scale information.

Although 2D methods work reasonably well, they do not consider volume images as a whole since they only analyze the image data slice by slice, neglecting information from other slices. U-Net-inspired 3D networks [7, 8, 27] are more effective than 2D U-Net on 3D volumes that have similar cross-voxel distances in all three dimensions and they have been applied to various tasks [6, 33, 40, 41, 44].

Transformer-based medical image segmentation models are increasing in popularity. They are usually based on CNNs also due to the fact that it is hard to train a pure Transformer model on limited data [5, 12, 31, 32, 38, 45].

In terms of attention, the squeeze-and-excitation (SE) block [14] was proposed to explicitly model the interdependencies between the channels of its convolutional features. Beyond channel-wise attention, the CBAM [43] models both the channel and spatial attention, which were combined into a 3D attention map in the bottleneck attention module (BAM) [30]. Self-attention, cross-attention, and their variants have also been heavily used, especially in vision Transformers [5, 9, 29, 32, 38]. However, limited research has addressed attention between slices, particularly the concept of using attention to mimic clinical decision-making. RsaNet [48] incorporates attention in all three directions in 3D networks instead of 2D networks. SAU-Net [50] learns attention between different slices of the final feature map but does not learn the whole semantic information. AFTer-U-Net [46] applies axial attention to fuse axial information across different slices, but a hyperparameter specifies how many neighboring slices to include in the attention. SATr [23] uses a 2D encoder to encode the slice of interest as well as the upper and lower slices and then fuses the encoded feature maps 3D convolution while calculating the attention map with a Transformer block, but a hyperparameter specifies the number of upper and lower slices. GT-Net [10] applies the attention mechanism in the skip connection at the bottleneck layer of a U-Net. Our CAT-Net [15] uses the cross-slice information at different scales of the deep networks, allowing the model to learn more comprehensive volumetric cross-slice information. Its performance is superior to the other 2.5D methods, but the model requires the training of a large number of parameters.

3. Defining 2.5D Segmentation

We denote $x_0 \in \mathbb{R}^{l \times c_0 \times h_0 \times w_0}$ as the input volumetric image, where l is the total number of slices in the volume, c_0 is the number of channels per slice, and h_0 and w_0 are the height and width of the inputs. In most cases, there is only one channel per slice; i.e., $c_0 = 1$, for a single modality. However, there may be multiple channels with multi-parametric MRI, such as T1 weighted imaging, T2 weighted imaging, apparent diffusion coefficient (ADC) map, etc., or for multi-modality imaging, such as combined MRI and Computed Tomography (CT). We also denote a pure 2D encoder as E and decoder as D.

Conventionally, researchers stack nearby slices as additional channels in the input to 2D segmentation networks and segment the middle slice, referring to this as 2.5D segmentation [11, 49]; i.e.,

$$y = D(E(F_{\text{cat}}(x))), \tag{1}$$

where y is the segmentation mask and F_{cat} denotes the concatenation of nearby slices. Methods that use other ways to combine the input images in the volume, such as those of Yu et al. [47] and Wang et al. [42], can also be expressed in the form of (1).

By contrast, recent models [10, 15, 23, 46, 48] incorporate different cross-slice attention mechanisms on deep feature maps between the encoder and decoder, expressible as

$$y = D(F_{\text{attention}}(E(x))), \tag{2}$$

where $F_{\text{attention}}$ is the cross-slice attention operation between the encoder and decoder. Most of them do not explicitly claim to be 2.5D methods, but we regard them as such since they consider the relationship between slices.

Additionally, SAU-Net [50] uses a cross-slice attention mechanism after the decoder to learn the cross-slice relationship, which can be expressed as

$$y = F_{\text{attention}}(D(E(x))), \tag{3}$$

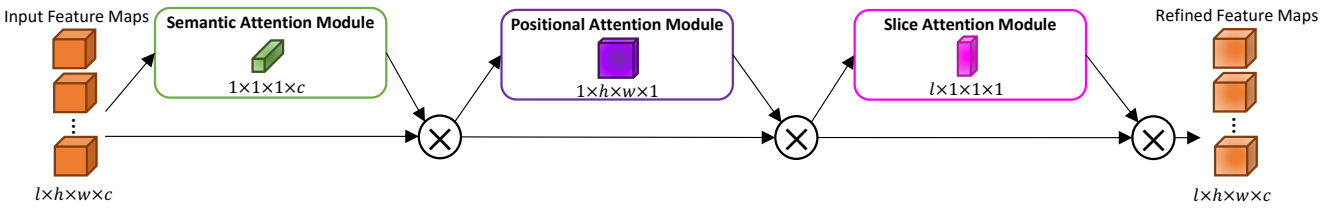

Figure 1. The CSAM architecture. The input feature maps from different slices are fed into the module concurrently, where they go through three sequential sub-modules: the semantic, positional, and slice attention modules.

where $F_{\text{attention}}$ is the cross-slice attention operation following the decoder.

Since there has been no formal consensus on the 2.5D segmentation approach, we define the general form of 2.5D segmentation as

$$y = F_{\text{post}}(D(F_{\text{mid}}(E(F_{\text{pre}}(x))))), \tag{4}$$

where F_{pre} is a function applied to the input volume, F_{mid} is the operation between the 2D encoder and decoder, and F_{post} is the operation after the decoder. Thus, *for any 2D encoder E and decoder D, when at least one of F_{pre}, F_{mid}, or F_{post} in (4) involves operations between different slices, we categorize the segmentation model as being 2.5D.* Note that, (4) reduces to (1) when F_{pre} represents a concatenation of nearby slices while $F_{\text{mid}} = F_{\text{post}} = I$ (i.e., identity functions), it reduces to (2) when F_{mid} represents cross-slice attention between the encoder and decoder while $F_{\text{pre}} = F_{\text{post}} = I$, and it reduces to (3) when F_{post} becomes the cross-slice attention mechanism while $F_{\text{pre}} = F_{\text{post}} = I$.

4. The Cross-Slice Attention Module

4.1. Overview

The CSAM incorporates information from other slices within the image volume to perform semantic, positional, and slice attention in 2.5D image segmentation. The CSAM inputs the feature maps from all slices at different scales and outputs the refined feature maps. As shown in Figure 1, given feature maps $F \in \mathbb{R}^{l\times h\times w\times c}$ from all the slices in the volume, where l is the total number of slices in the volume, h and w determine the size of the feature maps, and c is the number of channels, the CSAM sequentially calculates the semantic $M_{\text{semantic}} \in \mathbb{R}^{1\times 1\times 1\times c}$, positional $M_{\text{positional}} \in \mathbb{R}^{1\times h\times w\times 1}$, and slice $M_{\text{slice}} \in \mathbb{R}^{l\times 1\times 1\times 1}$ attention maps, which are used to weigh the input feature maps F, thus obtaining the refined feature maps F' as follows:

$$F_1 = M_{\text{semantic}}(F) \otimes F, \tag{5}$$
$$F_2 = M_{\text{positional}}(F_1) \otimes F_1, \tag{6}$$
$$F' = M_{\text{slice}}(F_2) \otimes F_2, \tag{7}$$

where $F_1, F_2 \in \mathbb{R}^{l\times h\times w\times c}$ are intermediate results and $\otimes$ denotes element-wise multiplication. We perform broadcasts accordingly during the multiplication.

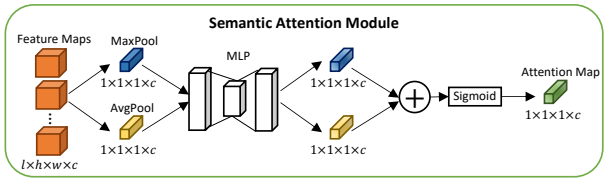

Figure 2. The semantic attention.

4.2. Semantic Attention

The semantic attention module (Figure 2) takes feature maps F as input and produces a semantic attention map M_{semantic}, which focuses on what semantic information is important among all the 2D feature maps of the volume. Unlike conventional 2D approaches, our semantic attention map is the same for all 2D feature maps in the volume. To compute the semantic feature map, the l, h, and w dimensions are squeezed by max and average pooling as these pooling methods gather different information about the features for better semantic attention [43]. First, global max and average pooling are performed on the l, h and w dimensions. Second, the max and averagepooled features are separately fed into the same multi-layer perceptron (MLP). Finally, the sum of the two outputs is fed into a sigmoid function to generate the semantic attention map. The operations can be written as

$$M_{\text{semantic}}(F) = \sigma(\text{MLP}(\text{MP}(F)) + \text{MLP}(\text{AP}(F))), \tag{8}$$

where $\sigma(\cdot)$ is the sigmoid function and $\text{MP}(\cdot)$ and $\text{AP}(\cdot)$ denote max and average pooling, respectively. Note that the $l \times h \times w \times c$ input feature maps are pooled to $1 \times 1 \times 1 \times c$; i.e., the semantic attention weights are the same on a single channel at different spatial locations within the slice and between slices.

4.3. Positional Attention

The positional attention module (Figure 3) focuses on where the important information is on each slice of the input volume. It takes feature maps F_1 as inputs and outputs a positional attention map $M_{\text{positional}}$. The features maps are max and averagepooled across l and c dimensions to two

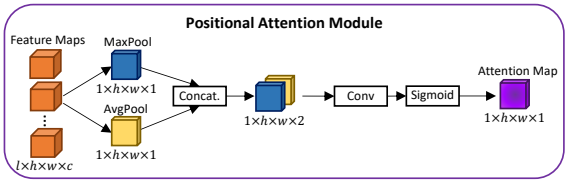

Figure 3. The positional attention.

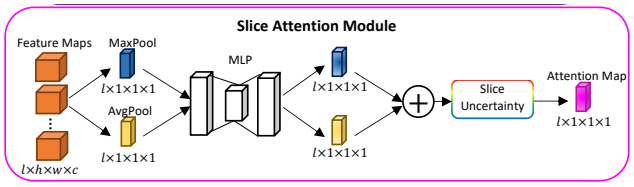

Figure 4. The slice attention.

$1 \times h \times w \times 1$ feature maps that are subsequently concatenated and then subjected to a convolution operation and a sigmoid function, as follows:

$$M_{\text{positional}}(F_1) = \sigma(\text{conv}[\text{MP}(F_1); \text{AP}(F_1)]), \quad (9)$$

where conv$[.;.]$ denotes a convolution operation with the semicolon denoting concatenation.

4.4. Slice Attention

The slice attention module (Figure 4) is designed to compute which slices are more likely to contain important information relevant to the intended task and suppress features that might lead to false positive predictions on particular slices; e.g., the module should infer that it is less likely to have positive prostate segmentation in the first and last slice of the volume than in the middle of the volume. Furthermore, the slice uncertainty block within the slice attention module can capture the aleatoric uncertainty inherently stemming from the input data while also acting as a form of regularization to improve performance.

The slice attention module first max and average pools the input feature maps and feeds them separately into the same MLP. However, here the input feature maps are pooled in h, w, and c dimensions, making the pooled features $l \times 1 \times 1 \times 1$. Before generating the slice attention map M_{slice}, the outputs of the MLP are added together and fed into a slice uncertainty block, which captures the uncertainty caused by the partial volume effects [2] or the difference in length between the slice thickness and the spatial resolution. The slice attention on different slices should be correlated, and the correlation should be stronger on nearby slices. Therefore, we use a low-rank Gaussian distribution $\mathcal{N}(\mu, \Sigma)$ to model the slice uncertainty so that the three components describing the distribution, namely the mean μ, covariance factor P, and the covariance diagonal D can all be computed efficiently while approximating the underlying Gaussian distribution. The input to the slice uncertainty block is $V \in \mathbb{R}^{l \times 1}$, which is the reshaped tensor of the summation of two $l \times 1 \times 1 \times 1$ tensors. Three linear operations are performed on this tensor:

$$\mu' = W_\mu V, \qquad P' = W_P V, \qquad D' = W_D V, \quad (10)$$

where $W_\mu \in \mathbb{R}^{l \times l}$, $W_D \in \mathbb{R}^{l \times l}$, $W_P \in \mathbb{R}^{(lr) \times l}$, and r is the rank of the parameterization. The covariance matrix Σ of the distribution is calculated as

$$\Sigma = PP^T + D, \quad (11)$$

where $P \in \mathbb{R}^{l \times r}$ is the reshaped tensor of P' and $D \in \mathbb{R}^{l \times l}$ is a diagonal matrix with elements of D' on the diagonal. The mean of the low-rank Gaussian distribution is $\mu \in \mathbb{R}^l$, which is the reshaped tensor of μ'. Next, a vector z is sampled from the distribution and passed through a sigmoid function to obtain the un-reshaped slice attention map

$$M'_{\text{slice}} = \sigma(z); \qquad z \sim \mathcal{N}(\mu, \Sigma). \quad (12)$$

The final output $M_{\text{slice}} \in \mathbb{R}^{l \times 1 \times 1 \times 1}$ is the reshaped M'_{slice}.

4.5. Implementation on Different Backbones

CSAMs are incorporated into U-Net-like backbone networks with skip connections between the encoder E and decoder D. We define $X_{i,j}$ as a basic convolution block of the backbone network and $x_{i,j}$ as the output of convolution block $X_{i,j}$. E is made up of the convolution blocks $X_{i,0}$, and all the other convolution blocks make up D.

Taking in a tensor $x \in \mathbb{R}^{l \times c_0 \times h_0 \times w_0}$ with l, c_0, h_0, and w_0 being the total number of slices in the volume, the number of channels, and the height and width of the inputs, respectively, E treats l as the batch dimension in conventional 2D approaches and outputs a set of encoded feature maps $\{x_{i,0}\}_{0 \le i < L} = E(x_0)$, where $x_{i,0} \in \mathbb{R}^{l \times c_i \times h_i \times w_i}$ is the output of each convolution block and L is the number of layers in the network.

We use D_i to represent layer i of the decoder and d_i as the corresponding output, where $0 \le i < L - 1$. We manually define the output of CSAM_{L-1}, which is the deepest CSAM, as d_{L-1}; i.e. $d_{L-1} = \text{CSAM}_{L-1}(x_{L-1,0})$. When $0 \le i < L - 1$, in a U-Net-based framework, D_i is equivalent to $X_{i,1}$, d_i is the same as $x_{i,1}$, and the final output is $x_{0,1}$, whereas in a U-Net++-based framework, D_i consists of a set of convolution blocks, namely $D_i = \{X_{i,j}\}_{1 \le j < L-i}$ and $d_i = \{x_{i,j}\}_{1 \le j < L-i}$, and the final output is $x_{0,L-1}$. As shown in Figure 5, the deep feature maps $x_{i,0}$ at each encoder layer i are injected into the CSAMs before getting to the decoder:

$$d_i = D_i(\text{CSAM}_i(x_{i,0}), \mathcal{U}(d_{i+1})), \quad (13)$$

where CSAM_i is associated with layer i, and $\mathcal{U}(.)$ denotes the upsampling operation.

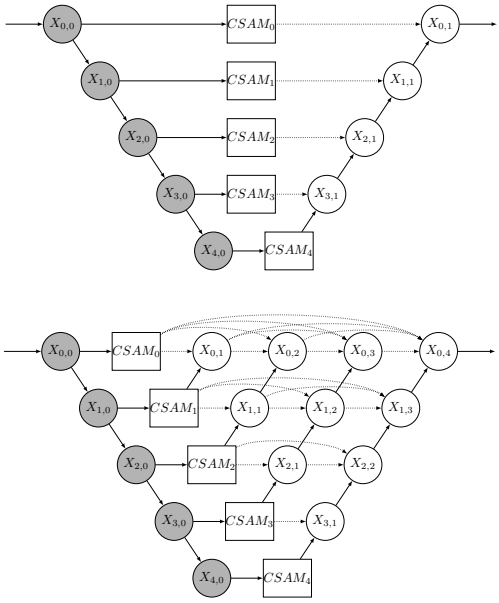

Figure 5. Implementation of the CSAM (top) in the nnU-Net [17] and MSU-Net [36], and (bottom) in the nnU-Net++ [51]. The gray circles represent the 2D encoder. The white circles represent the 2D decoder.

Semantic	Positional	Slice	TZ	PZ
			87.9***	82.4***
✓			88.0***	83.2***
	✓		88.0***	83.0***
		✓	88.3***	83.7***
✓	✓		88.7***	84.2***
✓		✓	88.5***	83.9***
	✓	✓	88.4***	84.0***
✓	✓	✓	**89.9**	**85.2**

Table 1. Ablation study of zonal segmentation (DSC in %). Note: in this and subsequent tables, the labels *, **, *** indicate p-values of ≤ 0.1, ≤ 0.05, and ≤ 0.01, respectively. All the experiments involving cross-validation adhere to the same notation.

	TZ	PZ
Slice without uncertainty	87.6***	82.8***
Slice with uncertainty	**88.3**	**83.7**
All without uncertainty	88.3***	84.7***
All with uncertainty	**89.9**	**85.2**

Table 2. Ablation study of the effect of using slice uncertainty (DSC in %).

5. Experiments

5.1. Implementation and Evaluation Details

The CSAM was tested on segmentation of prostate, placenta, and cardiac MRI, with nnU-Net [17], MSU-Net [36], and nnU-Net++ [51] as backbones. It was also evaluated on prostate MRI cancer segmentation. Cross entropy loss was used on prostate zonal and cardiac segmentation, dice loss for placenta segmentation, and focal loss [24] for prostate cancer (PCa) segmentation. We used the Adam [22] optimizer with a learning rate of 0.0001 and weight decay regularization with the parameter set to 1×10^{-5}. Only center crop, horizontal flip, and Gamma transform were performed as augmentation.

The Dice similarity coefficient (DSC) and relative absolute volume difference (RAVD) were calculated in 3D. Patient-wise classification AUC was calculated for prostate MRI cancer segmentation. For experiments involving cross-validation, we performed statistical analysis with the Mann-Whitney U Test [28] to compare the distribution of results from our CSAM-based models against those of the competing models.

5.2. Prostate MRI Zonal Segmentation

This dataset comprises 296 patients, including only T2-weighted MRI scans with an in-plane resolution of 0.625 mm^2 and a through-plane resolution of 3 mm. The prostate transition zone (TZ) and peripheral zone (PZ) were manually annotated by clinical experts as ground truth. We randomized the data and grouped them into five folds for cross-validation. For the training of the 2D and 2.5D models, the input volume size is $128 \times 128 \times 20$. To perform multiple downsampling operations in the 3D models, we upsampled the volumes along the z-axis to $128 \times 128 \times 80$. We carried out 5-fold cross-validation in this experiment.

Our ablation study was performed on the prostate zonal segmentation using a nnU-Net++-based architecture. As reported in Table 1, all three sub-modules can improve the segmentation performance, while using attention in all three dimensions achieves the best results with statistical significance in terms of DSC in both the TZ and PZ. Furthermore, we compared model performance including and excluding the slice uncertainty on the same task. From Table 2, we see that compared to the absence of slice uncertainty modeling, both using just the slice attention module and using all three attention modules along with the slice uncertainty can improve performance by a good amount.

To test the generalizability of our CSAM across different backbones, we incorporated it into nnU-Net, MSU-Net, and nnU-Net++, and compared them with the corresponding 2D model, 3D model, and 2.5D GT-Net [10], SA-Net [50], Satr-Net [23], and CAT-Net [15]. Table 3 shows that on nnU-Net, CSAMs significantly increased segmentation performance relative to the 2D and 3D nnU-Net for both the TZ and PZ, while the performance is similar for the TZ compared with

	TZ		PZ		
	DSC (%)↑	RAVD (%)↓	DSC (%)↑	RAVD (%)↓	# Parameters
nnU-Net [17]	88.5***	23.6***	83.6***	32.8***	138M
3D nnU-Net [17]	89.0***	22.8***	83.6***	32.1***	264M
GT-nnU-Net [10]	87.7***	25.6***	82.7***	34.6***	186M
SAnnU-Net [50]	87.6***	25.3***	82.3***	35.4***	138M
Satr-nnU-Net [23]	85.8***	29.8***	79.0***	43.2***	269M
CAT-nnU-Net [15]	89.5	**20.8**	85.1**	29.8**	614M
CSAM-nnU-Net	**89.7**	21.0	**85.7**	**29.1**	139M
MSU-Net [36]	88.1	24.8**	82.3	33.5*	47.1M
3D MSU-Net [36]	**89.3**	**21.8**	**83.6**	32.6	293M
GT-MSU-Net [10]	86.1***	28.5***	79.2***	42.0***	55.0M
SAMSU-Net [50]	87.1**	26.5**	81.1**	37.6**	47.1M
Satr-MSU-Net [23]	84.4***	31.8***	78.8***	41.2***	87.1M
CAT-MSU-Net [15]	87.0***	25.3***	81.7*	34.0**	264M
CSAM-MSU-Net	89.0	**21.8**	83.1	**31.8**	47.2M
nnU-Net++ [51]	87.9***	23.9***	82.4***	34.0**	36.6M
3D nnU-Net++ [51]	88.1***	23.8***	82.2***	34.2***	110M
GT-nnU-Net++ [10]	88.6***	23.3***	83.8**	33.0**	57.6M
SAnnU-Net++ [50]	85.1***	30.6***	77.7***	44.5***	36.6M
Satr-nnU-Net++ [23]	88.4***	24.1***	83.3***	33.9***	101M
CAT-nnU-Net++ [15]	**89.9**	**20.6**	**85.6**	**29.1**	398M
CSAM-nnU-Net++	**89.9**	**20.6**	85.2	29.5	36.8M

Table 3. Comparison of CSAM-Net against the corresponding 2D, 3D, and 2.5D models for prostate zonal segmentation.

CAT-nnU-Net. However, CSAM-nnU-Net significantly outperforms CAT-nnU-Net for the segmentation of the PZ, while employing considerably fewer parameters. The other 2.5D models underperformed the 2D nnU-Net model due to the fact that prostate zonal segmentation requires relatively little volumetric information and less carefully designed attention mechanisms can confuse the model. In terms of MSU-Net-based network structures, CSAM-MSU-Net significantly outperforms 2D MSU-Net, GT-MSU-Net, SAMSU-NET, Satr-MSU-Net, and CAT-MSU-Net, while exhibiting similar performance to 3D MSU-Net. In nnU-Net++, CSAMs have similar performance to CAT modules, while outperforming the corresponding 2D and 3D methods. In short, CSAM-Net showed superior performance when using any of nnU-Net, MSU-Net, and nnU-Net++ architectures in prostate zonal segmentation. Furthermore, the performance of CSAM-Net is similar to that of CAT-Net or the corresponding 3D method, but it remains a favorable option due to its much smaller number of parameters.

We evaluated the slice uncertainty with CSAM-nnU-Net. We investigated the lower quartile, median, and upper quartile of the uncertainty values of the correctly and incorrectly segmented pixels. The upper quartile uncertainty value of correctly segmented pixels was 0, while the incorrectly segmented pixels had a lower quartile of 0.04, a median of 0.09, and an upper quartile of 0.15. This indicates that uncertainty estimation can be used to discriminate between correctly and incorrectly segmented pixels. Furthermore, by examining the pixel-wise segmentation error rates corresponding to each uncertainty value, we found that the correlation between the error rates and the uncertainty values is $r = 0.791$.

5.3. Prostate MRI Cancer Segmentation

The dataset consists of 652 patients with T2-weighted imaging, ADC, and high b-value (b=1400) diffusion-weighted imaging (DWI) MRI, with volume sizes of $128 \times 128 \times 20$. The in-plane resolution is $0.625\,\text{mm}^2$, and the through-plane resolution is 3 mm. They were stacked together with the manual zonal segmentation as inputs to the models. 220 patients were confirmed with PCa, while 432 patients had no indication of PCa with biopsy. 5-fold cross-validation was applied in the experiment.

We evaluated the performance of the different models by DSC and the classification AUC. The cancer lesion is small relative to the entire image and, therefore, a DSC value for a cancer lesion is expected to be lower than for zonal segmentation. Since 2/3 of the data are patients without cancer, we also measured the classification AUC. Regarding classification, we define that only segmentations with no segmented cancer are negative patients. We compared

	DSC (%)↑	cls. AUC ↑
3D nnU-Net [17]	37.0**	0.854**
3D Residual U-Net [4]	39.3**	0.868**
3D SEResU-Net [14]	45.1	0.850***
3D Attention U-Net [29]	45.5	0.841***
VNet [27]	44.1*	0.846**
VoxResNet [7]	47.2	0.868**
UNETR [12]	43.6	0.852**
VTU-Net [31]	44.3*	0.846***
CSAM-nnU-Net	**51.0**	**0.942**

Table 4. Results of prostate cancer segmentation.

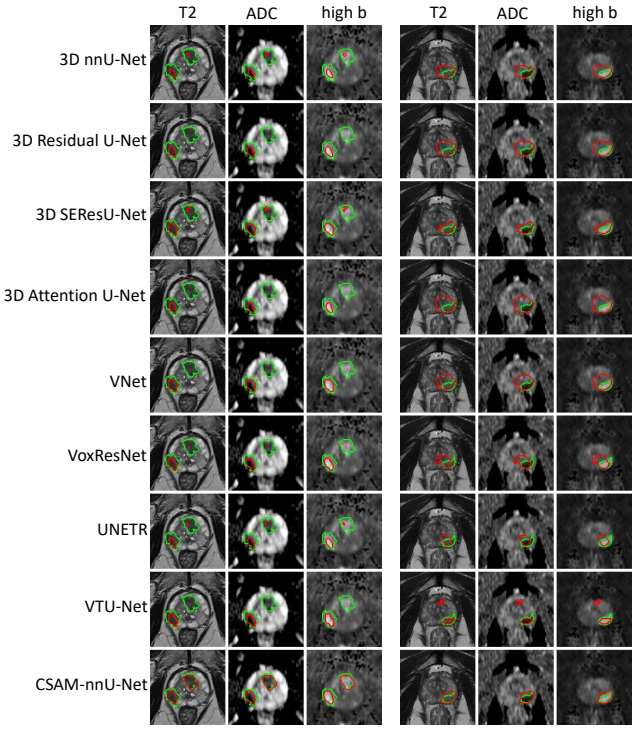

Figure 6. Two qualitative results of PCa segmentation. Green indicates the ground truth and red indicates the segmentation results by the corresponding models.

our CSAM-nnU-Net with other state-of-the-art 3D segmentation models, 3D nnU-Net [17], 3D Residual U-Net [4], 3D SEResNet [14], 3D Attention U-Net [29], VNet [27], VoxResNet [7], UNETR [12], and VTU-Net [31], and the results are reported in Table 4. CSAM-nnU-Net performed the best in both segmentation and classification. The qualitative results are shown in Figure 6, which reveals that only CSAM-nnU-Net is able to segment both lesions in the left example, whereas other models all failed to segment the lesion on the right. Additionally, since the lesions appear to be slightly different among T2, ADC, and high-b images, the manual ground truth label fits the T2 image better while the segmentation by CSAM-nnU-Net fits the other images better. In the example on the right, other models either over-predicted the segmentation or generated segmentations with wrong lesion shapes, while the CSAM-nnU-Net result was closest to the ground truth.

	DSC (%)↑	RAVD (%)↓
nnU-Net [17]	81.2	36.5
3D nnU-Net [17]	64.1	49.8
GT-nnU-Net [10]	82.0	34.8
SAnnU-Net [50]	82.6	34.7
CSAM-nnU-Net	**83.8**	**32.5**
MSU-Net [36]	65.5	51.8
3D MSU-Net [36]	62.3	55.0
GT-MSU-Net [10]	63.6	54.9
SAMSU-Net [50]	**69.4**	53.7
CSAM-MSU-Net	**69.4**	**50.9**
nnU-Net++ [51]	65.6	53.7
3D nnU-Net++ [51]	70.7	46.4
GT-nnU-Net++ [10]	83.0	**34.2**
SAnnU-Net++ [50]	62.8	52.9
CSAM-nnU-Net++	**83.2**	34.7

Table 5. Placenta segmentation results (no statistical testing).

5.4. Placenta MRI Segmentation

The data was acquired in three orthogonal planes (axial, sagittal, and coronal), including 14 to 18 weeks gestational age (GA), from eligible pregnant women [19]. It includes 150 pregnancies, among which 105, 15, and 30 were used for training, validation, and testing, respectively. For training, we resized the volumes to $128 \times 128 \times 64$ for all three imaging views.

We compared the CSAM-based models with the corresponding 2D and 3D models, as well as GT-based [10] and SA-based [50] 2.5D models on nnU-Net, MSU-Net, and nnU-Net++. As shown in Table 5, our CSAM-based models are always the ones with the best performance, while other modules or methods are inconsistent on different backbones. For example, SAnnU-Net and SAMSU-Net are not too much worse than the corresponding CSAM-based models, but SAnnU-Net++ is the worst among all the competing nnU-Net++-based models. The GT method works fine on nnU-Net and nnU-Net++, but it is worse than the pure 2D model when it is applied on MSU-Net. An example of the segmentation by nnU-Net++-based models is shown in Figure 7, and we see that nnU-Net++, 3D nnU-Net++, and SAnnU-Net++ under-segment the placenta, while GT-nnU-Net++ fails to segment the placenta on the first slice. Our CSAM-based method yields the best result.

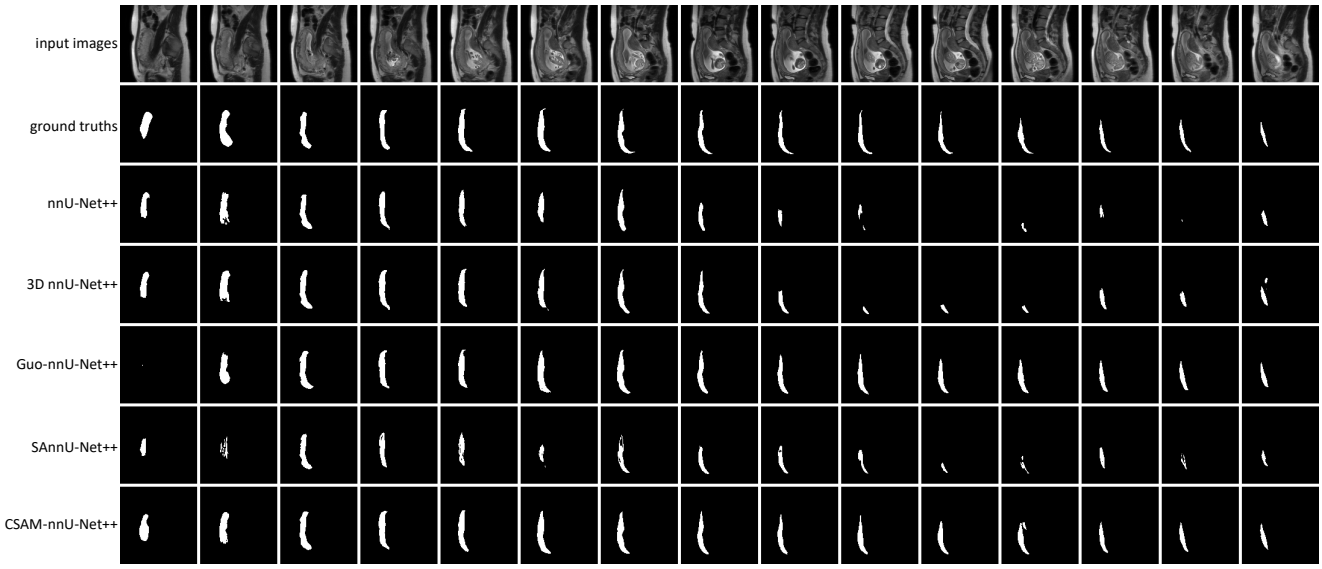

Figure 7. Results of placenta segmentation on consecutive slices. All the other nnU-Net++-based methods under-segment.

5.5. Cardiac MRI Segmentation

We utilized the ACDC public dataset [3], which consists of 100 subjects (20 healthy subjects, 20 subjects with previous myocardial infarction, 20 subjects with dilated cardiomyopathy, 20 subjects with hypertrophic cardiomyopathy, and 20 subjects with abnormal right ventricle). The segmentation labels included right ventricular endocardium (RV), myocardium (MYO), and left ventricular endocardium (LV) for both the end diastole and end systole phases. We performed 5-fold cross-validation in this experiment.

The results are reported in Table 6. Our CSAM showed a consistent improvement over the corresponding 2D and 2.5D models, and the CSAM-based models were always the best-performing methods. Even in some cases where the CSAM-based methods do not have the highest DSC, there is no statistical difference between the performance of the CSAM-based methods and the competing methods.

	RV	MYO	LV	Average
nnU-Net [17]	**86.8**	**85.5**	92.4	89.2
GT-nnU-Net [10]	86.3	**85.5**	**92.5**	89.2
SAnnU-Net [50]	86.2	85.1	92.2	88.9**
CAT-nnU-Net [15]	84.8*	84.0**	91.8	88.1**
CSAM-nnU-Net	**86.8**	**85.5**	**92.5**	**89.3**
MSU-Net [36]	82.9	83.6**	91.3	87.0*
GT-MSU-Net [10]	**83.8**	83.6*	91.2	87.4
SAMSU-Net [50]	82.4	82.4***	90.6**	86.3*
CAT-MSU-Net [15]	79.3**	81.3***	90.1**	85.3**
CSAM-MSU-Net	83.3	**84.3**	**91.8**	**87.7**
nnU-Net++ [51]	83.0**	83.0***	90.3***	86.5***
GT-nnU-Net++ [10]	86.5	**85.7**	92.5	**89.2**
SAnnU-Net++ [50]	82.1**	81.6***	89.8***	85.6***
CAT-nnU-Net++ [15]	86.2	84.7**	92.1*	88.9*
CSAM-nnU-Net++	**86.6**	85.4	**92.7**	**89.2**

Table 6. Comparison on cardiac MRI segmentation (DSC in %).

6. Conclusions

For the segmentation for anisotropic volumetric MRI data, we have devised a 2.5D cross-slice attention module (CSAM) that introduces minimal additional parameters to train, which can be incorporated into a variety of segmentation network backbones. Our extensive experiments showed that, despite the relatively small number of additional parameters, our CSAM-based models outperform the associated baseline 2D, 3D, and 2.5D models under simple training and data augmentation settings consistently across different segmentation tasks. Our study confirms that the CSAM's ability to learn and leverage cross-slice information within 3D image volumes improves volumetric segmentation performance over 2D and 3D methods and that our models generalize better than the state-of-the-art 2.5D modules.

Our research opens up promising avenues for future work. In particular, since pooling is performed globally in CSAM, it may be possible to improve performance and generalizability by incorporating into the cross-slice attention strategy information that is more local. Additionally, a further study of uncertainty measurement is needed, as this is critical to understanding the limitations of deep learning-based medical image segmentation models, which would help clinicians make better informed decisions.

References

[1] M. Bakator and D. Radosav, "Deep learning and medical diagnosis: A review of literature," *Multimodal Technologies and Interaction*, vol. 2, no. 3, p. 47, 2018. 1

[2] M. A. G. Ballester, A. P. Zisserman, and M. Brady, "Estimation of the partial volume effect in MRI," *Medical Image Analysis*, vol. 6, no. 4, pp. 389–405, 2002. 4

[3] O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, X. Yang, P.-A. Heng, I. Cetin, K. Lekadir, O. Camara, M. A. G. Ballester *et al.*, "Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved?" *IEEE Transactions on Medical Imaging*, vol. 37, no. 11, pp. 2514–2525, 2018. 8

[4] M. Bhalerao and S. Thakur, "Brain tumor segmentation based on 3D residual U-Net," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, 2020, pp. 218–225. 7

[5] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-Unet: Unet-like pure transformer for medical image segmentation," *arXiv preprint arXiv:2105.05537*, 2021. 2

[6] J. Chang, X. Zhang, M. Ye, D. Huang, P. Wang, and C. Yao, "Brain tumor segmentation based on 3D Unet with multi-class focal loss," in *11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, 2018, pp. 1–5. 2

[7] H. Chen, Q. Dou, L. Yu, J. Qin, and P.-A. Heng, "VoxResNet: Deep voxelwise residual networks for brain segmentation from 3d mr images," *NeuroImage*, vol. 170, pp. 446–455, 2018. 1, 2, 7

[8] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning dense volumetric segmentation from sparse annotation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2016, pp. 424–432. 2

[9] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020. 2

[10] D. Guo and D. Terzopoulos, "A transformer-based network for anisotropic 3D medical image segmentation," in *25th International Conference on Pattern Recognition (ICPR)*, 2021, pp. 8857–8861. 1, 2, 5, 6, 7, 8

[11] L. Han, Y. Chen, J. Li, B. Zhong, Y. Lei, and M. Sun, "Liver segmentation with 2.5 D perpendicular UNets," *Computers & Electrical Engineering*, vol. 91, p. 107118, 2021. 1, 2

[12] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu, "UNETR: Transformers for 3D medical image segmentation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 574–584. 2, 7

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778. 2

[14] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141. 2, 7

[15] A. L. Y. Hung, H. Zheng, Q. Miao, S. S. Raman, D. Terzopoulos, and K. Sung, "CAT-Net: A cross-slice attention transformer model for prostate zonal segmentation in MRI," *IEEE Transactions on Medical Imaging*, vol. 42, no. 1, pp. 291–303, 2022. 1, 2, 5, 6, 8

[16] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning*, 2015, pp. 448–456. 2

[17] F. Isensee, J. Petersen, A. Klein, D. Zimmerer, P. F. Jaeger, S. Kohl, J. Wasserthal, G. Koehler, T. Norajitra, S. Wirkert *et al.*, "nnU-Net: Self-adapting framework for U-Net-based medical image segmentation," *arXiv preprint arXiv:1809.10486*, 2018. 1, 2, 5, 6, 7, 8

[18] F. Isensee, P. F. Jaeger, P. M. Full, I. Wolf, S. Engelhardt, and M.-H. K. H., "Automatic cardiac disease assessment on cine-MRI via time-series segmentation and domain specific features," in *International Workshop on Statistical Atlases and Computational Models of the Heart*, 2017, pp. 120–129. 1

[19] C. Janzen, M. Y. Lei, B. R. Lee, S. Vangala, I. DelRosario, Q. Meng, B. Ritz, J. Liu, M. Jerrett, T. Chanlaw *et al.*, "A description of the imaging innovations for placental assessment in response to environmental pollution study (PARENTs)," *American Journal of Perinatology*, no. AAM, 2022. 7

[20] H. Jia, Y. Xia, Y. Song, D. Zhang, H. Huang, Y. Zhang, and W. Cai, "3D APA-Net: 3D adversarial pyramid anisotropic convolutional network for prostate segmentation in MR images," *IEEE Transactions on Medical Imaging*, vol. 39, no. 2, pp. 447–457, 2019. 1

[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," *Biocybernetics and Biomedical Engineering*, vol. 40, no. 3, pp. 1314–1327, 2020. 2

[22] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014. 5

[23] H. Li, L. Chen, H. Han, and S. Kevin Zhou, "SATr: Slice attention with transformer for universal lesion detection," in *Proceedings of the Medical Image Computing and Computer Assisted Intervention Conference*, 2022, pp. 163–174. 1, 2, 5, 6

[24] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 02, pp. 318–327, 2020. 5

[25] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017. 1

[26] Y. Liu, F. Zabihollahy, R. Yan, B. Lee, C. Janzen, S. U. Devaskar, and K. Sung, "Evaluation of spatial attentive deep learning for automatic placental segmentation on longitudinal MRI," *Journal of Magnetic Resonance Imaging*, 2022. 1

[27] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *Fourth International Conference on 3D Vision (3DV)*, 2016, pp. 565–571. 1, 2, 7

[28] N. Nachar *et al.*, "The Mann-Whitney U: A test for assessing whether two independent samples come from the same distribution," *Tutorials in Quantitative Methods for Psychology*, vol. 4, no. 1, pp. 13–20, 2008. 5

[29] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert, "Attention U-Net: Learning where to look for the pancreas," in *Medical Imaging with Deep Learning*, 2018. 2, 7

[30] J. Park, S. Woo, J.-Y. Lee, and I. S. Kweon, "BAM: Bottleneck attention module," *arXiv preprint arXiv:1807.06514*, 2018. 2

[31] H. Peiris, M. Hayat, Z. Chen, G. Egan, and M. Harandi, "A robust volumetric transformer for accurate 3D tumor segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2022, pp. 162–172. 2, 7

[32] 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 *International Workshop on Machine Learning in Medical Imaging*, 2021, pp. 267–276. 2

[33] S. Qamar, H. Jin, R. Zheng, P. Ahmad, and M. Usama, "A variant form of 3D-UNet for infant brain segmentation," *Future Generation Computer Systems*, vol. 108, pp. 613–623, 2020. 2

[34] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015, pp. 234–241. 1, 2

[35] L. Rundo, C. Han, Y. Nagano, J. Zhang, R. Hataya, C. Militello, A. Tangherloni, M. S. Nobile, C. Ferretti, D. Besozzi *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. 2

[36] R. Su, D. Zhang, J. Liu, and C. Cheng, "MSU-Net: Multi-scale U-Net for 2D medical image segmentation," *Frontiers in Genetics*, vol. 12, p. 140, 2021. 1, 2, 5, 6, 7, 8

[37] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," *arXiv preprint arXiv:1607.08022*, 2016. 2

[38] J. M. J. Valanarasu, P. Oza, I. Hacihaliloglu, and V. M. Patel, "Medical transformer: Gated axial-attention for medical image segmentation," *arXiv preprint arXiv:2102.10662*, 2021. 2

[39] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008. 2

[40] B. Wang, Y. Lei, S. Tian, T. Wang, Y. Liu, P. Patel, A. B. Jani, H. Mao, W. J. Curran, T. Liu *et al.*, "Deeply supervised 3D fully convolutional networks with group dilated convolution for automatic MRI prostate segmentation," *Medical Physics*, vol. 46, no. 4, pp. 1707–1718, 2019. 2

[41] 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," *arXiv preprint arXiv:1804.04341*, 2018. 2

[42] X. Wang, S. Han, Y. Chen, D. Gao, and N. Vasconcelos, "Volumetric attention for 3D medical image segmentation and detection," in *Proceedings of the Medical Image Computing and Computer Assisted Intervention Conference*, 2019, pp. 175–184. 1, 2

[43] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 3–19. 2, 3

[44] 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, 2020. 2

[45] G. Xu, X. Wu, X. Zhang, and X. He, "LeViT-UNet: Make faster encoders with transformer for medical image segmentation," *arXiv preprint arXiv:2107.08623*, 2021. 2

[46] X. Yan, H. Tang, S. Sun, H. Ma, D. Kong, and X. Xie, "AFTer-UNet: Axial fusion transformer UNet for medical image segmentation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 3971–3981. 1, 2

[47] Q. Yu, Y. Xia, L. Xie, E. K. Fishman, and A. L. Yuille, "Thickened 2D networks for efficient 3D medical image segmentation," *arXiv preprint arXiv:1904.01150*, 2019. 1, 2

[48] H. Zhang, J. Zhang, Q. Zhang, J. Kim, S. Zhang, S. A. Gauthier, P. Spincemaille, T. D. Nguyen, M. Sabuncu, and Y. Wang, "RSANet: Recurrent slice-wise attention network for multiple sclerosis lesion segmentation," in *Proceedings of the Medical Image Computing and Computer Assisted Intervention Conference*, 2019, pp. 411–419. 1, 2

[49] H. Zhang, A. M. Valcarcel, R. Bakshi, R. Chu, F. Bagnato, R. T. Shinohara, K. Hett, and I. Oguz, "Multiple sclerosis lesion segmentation with tiramisu and 2.5D stacked slices," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2019, pp. 338–346. 1, 2

[50] Y. Zhang, L. Yuan, Y. Wang, and J. Zhang, "SAU-Net: efficient 3D spine MRI segmentation using inter-slice attention," in *Medical Imaging With Deep Learning*, 2020, pp. 903–913. 1, 2, 5, 6, 7, 8

[51] 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*. Springer, 2018, pp. 3–11. 1, 5, 6, 7, 8