# NaMa: Neighbor-Aware Multi-Modal Adaptive Learning for Prostate Tumor Segmentation on Anisotropic MR Images

**Runqi Meng[1], Xiao Zhang[2,1], Shijie Huang[1], Yuning Gu[1], Guiqin Liu[5], Guangyu Wu[5], Nizhuan Wang[1], Kaicong Sun[1], Dinggang Shen[1,3,4] ***

[1]School of Biomedical Engineering & State Key Laboratory of Advanced Medical Materials and Devices, ShanghaiTech University
[2]School of Information Science and Technology, Northwest University
[3]Shanghai United Imaging Intelligence Co., Ltd.
[4]Shanghai Clinical Research and Trial Center
[5]Department of Radiology, Renji Hospital, Shanghai Jiao Tong University School of Medicine

## Abstract

Accurate segmentation of prostate tumors from multi-modal magnetic resonance (MR) images is crucial for diagnosis and treatment of prostate cancer. However, the robustness of existing segmentation methods is limited, mainly because these methods 1) fail to adaptively assess subject-specific information of each MR modality for accurate tumor delineation, and 2) lack effective utilization of inter-slice information across thick slices in MR images to segment tumor as a whole 3D volume. In this work, we propose a two-stage neighbor-aware multi-modal adaptive learning network (NaMa) for accurate prostate tumor segmentation from multi-modal anisotropic MR images. In particular, in the first stage, we apply subject-specific multi-modal fusion in each slice by developing a novel modality-informativeness adaptive learning (MIAL) module for selecting and adaptively fusing informative representation of each modality based on inter-modality correlations. In the second stage, we exploit inter-slice feature correlations to derive volumetric tumor segmentation. Specifically, we first use a Unet variant with sequence layers to coarsely capture slice relationship at a global scale, and further generate an activation map for each slice. Then, we introduce an activation mapping guidance (AMG) module to refine slice-wise representation (via information from adjacent slices) for consistent tumor segmentation across neighboring slices. Besides, during the network training, we further apply a random mask strategy to each MR modality to improve feature representation efficiency. Experiments on both in-house and public (PICAI) multi-modal prostate tumor datasets show that our proposed NaMa performs better than state-of-the-art methods.

## Introduction

Prostate cancer (PCa) is the leading cause of cancer-related death in men worldwide, and early detection of PCa is crucial for enhancing survival rates and providing appropriate intervention (Velonas et al. 2013; Boettcher et al. 2019). Clinical screening of PCa often uses multi-modal magnetic resonance imaging (MRI), including T2-weighted images (T2WIs), diffusion-weighted images (DWIs), and apparent diffusion coefficient (ADC) maps (derived from DWIs) (Tanimoto et al. 2007; Cornud et al. 2012). If a tumor is detected in the multi-modal MR images, its segmentation from surrounding normal tissues is required for cancer classification and treatment planning.

However, manual segmentation of tumors is time-consuming, and segmentation results are highly dependent on the expertise of the radiologists. Therefore, great amount

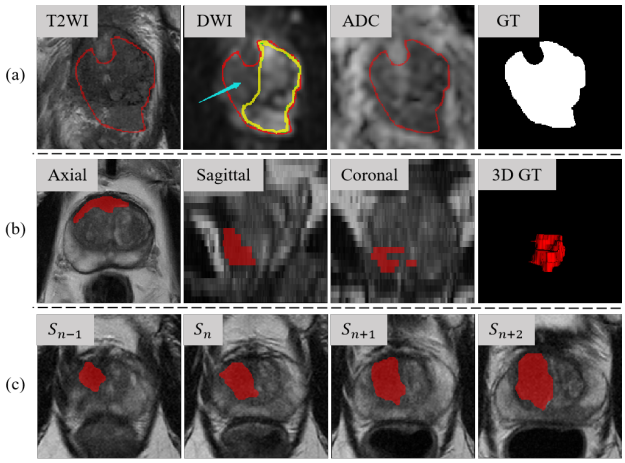


Figure 1: Characteristics of the multi-modal MR images used for prostate tumor segmentation. The red contours and masks represent the ground truth of the prostate tumor. (a) Some modalities may provide ineffective features, i.e., DWI image of one case has inhomogeneous intensities within the tumor (indicated by the blue arrow), which causes the under-segmentation issue (shown by yellow contour). (b) The MR images have high in-plane image resolution (axial view) and low through-plane resolution (vertical axis in sagittal and coronal views). (c) The locat ions of tumor across adjacent slices are inter-related.

*Corresponding author: Dinggang Shen (Dinggang.Shen@gmail.com).

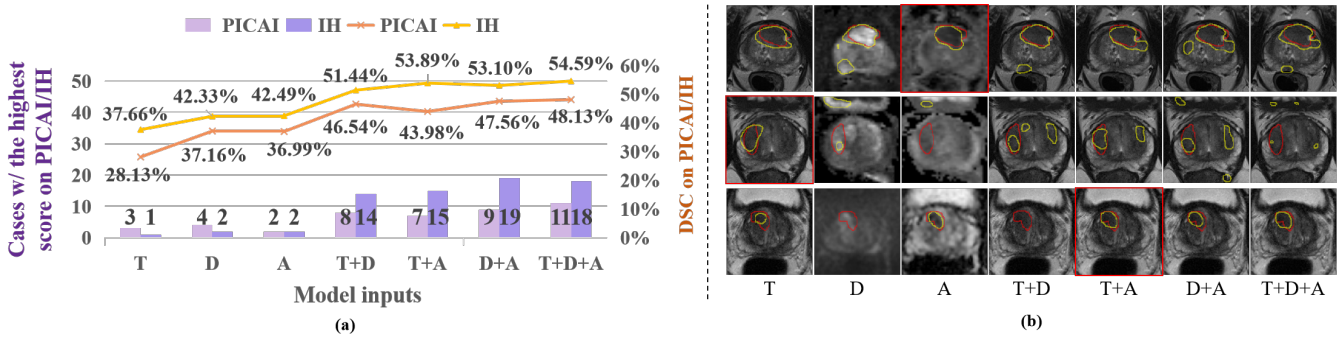


Figure 2: Analysis on the effectiveness of multi-modal fusion. The segmentation network uses a multi-branch encoder for multi-modal feature extraction, and a decoder to derive segmentation results using concatenated features from input modalities. 'T' denotes T2WI, 'D' denotes DWI, and 'A' denotes ADC map. (a) The line graph illustrates the Dice similariy coefficient (DSC) achieved by different input modalities on PICAI/IH. The histogram displays the distribution of the cases achieving optimal performance with each input modalities on PICAI/IH. (b) Visualized segmentation results for three cases using different input modalities, with each row for one case. The red contours indicate ground truth, and the yellow contours show segmentation output. The the input modality with the best performance in each case is highlighted by the red boxes. The results of 'T','T+D','T+A' and 'T+D+A' are overlaid on T2WIs, the results of 'D' and 'D+A' are overlaid on DWIs, and the results of 'A' are overlaid on ADC maps.

of efforts have been devoted to developing automated tumor segmentation methods. Although in recent years many deep learning-based methods have been proposed for tumor segmentation from multi-modal MR images, existing methods are usually limited in handling 1) effective fusion of multi-modal MR images, and 2) semantically consistent segmentation of 3D tumor across thick MR slices. All these limitations will be explained below.

With regard to multi-modal fusion, many works employ early, intermediate, or late fusion scheme (Pereira et al. 2016; Havaei et al. 2015) to integrate model-specific information for improved segmentation. However, the segmentation performances of such methods are usually affected by 1) unpredictable changes of image quality in the involved imaging modalities among different subjects, and 2) over-dependency of the segmentation model on certain modalities. These limitations are clearly revealed in our preliminary studies performed on both the in-house (IH) dataset and public dataset from PI-CAI2022 challenge (PICAI). First, as shown in Fig. 1(a), we illustrate the ineffectiveness of using a DWI image (with image noise and artifacts) cause inaccurate tumor segmentation. Actually, we find that only 18 cases (5.07%) in IH dataset and 11 cases (5.12%) in PICAI dataset possess superior performance when using all modalities, while other instances achieve good outcomes with only one or two modalities as depicted in Fig. 2(a). Second, we find that the segmentation results are biased towards the features from DWI and ADC maps as shown in Fig. 2(b), while the T2WIs usually contribute less to the segmentation results. However, there exist cases showing that results based on T2WIs agree better with the ground truth, with one example shown in the second row of Fig. 2(b). Therefore, more advanced segmentation method is required to adaptively find more reliable modalities for multi-modal segmentation.

Besides multi-modal fusion, semantically-consistent segmentation of tumors across thick slices is also critical for high-quality tumor segmentation. Several studies (Li et al. 2021a; Chen et al. 2020) have developed 3D tumor segmentation methods by modifying existing methods originally proposed for organ segmentation. However, existing 3D segmentation networks are typically designed for nearly-isotropic 3D images, and their application to clinical prostate MR images with thick slices usually results in limited performance (Zhang et al. 2020a) as shown in Fig. 1(b). Since MR images are mostly anisotropic, 2D CNN is often used in the current mainstream for tumor segmentation. Although some 2D CNN based methods leverage global or local inter-slice correlations to capture the spatial context (*i.e.,* 2.5D strategy), these 2.5D methods still lack effective feature aggregation strategy when dealing with complex anatomical variations and irregular tumor shapes.

To tackle the aforementioned challenges, in this study, we present a novel neighbor-aware multi-modal adaptive learning network (NaMa) to accurately segment prostate tumors from multi-modal anisotropic MR images. Specifically, we design a modality-informativeness adaptive learning module (MIAL) to implicitly assess the reliability of features in each modality, and adaptively fuse reliable representations across different modalities. Additionally, an activation mapping guidance module (AMG) is proposed to enable stable and consistent segmentation performance across slices in 3D space. Finally, to more effectively learn feature representations of different modalities, we apply the random mask strategy to the features of each modality in the network training stage. In summary, our main contributions are threefold:

- We propose a multi-modal adaptive learning strategy for modeling modal-specific informativeness to enable individual-effective multi-modal fusion.
- We propose an activation map guidance module to capture inter-slice relationship in anisotropic MR images to produce semantically consistent volumetric result.

- Extensive experiments on both in-house and public prostate tumor MRI datasets demonstrate that our model achieves significant improvement over many state-of-the-art methods.

# Related Work

## Multi-modal Fusion

In medical image analysis, accurate disease diagnoses could be achieved by exploring complementary information from multiple modalities. Early works distinguished fusion approaches into early fusion, intermediate fusion and late fusion, depending on where fusion is performed in the model (Zhang et al. 2020c). As early fusion suppresses intra-modal interactions and late fusion only aggregates the outputs from each modality without effectively capturing inter-modal interactions, the intermediate fusion strategy is the most prevalent in multi-modal learning, which can effectively model both intra- and inter-modality interactions.

To obtain more effective multi-modal representations, it is important to exploit cross-modal complementary information while mitigating modality-specific noises. Several methods (Zhan et al. 2021; Xing et al. 2022) introduce a gate mergence mechanism to automatically learn the weights of different modalities to enhance the task-related information. Additionally, several other methods (Zhang et al. 2022b; Jiale et al. 2023; Man, Gui, and Wang 2023) explore the complementary information from intra- and inter-modality by adopting various attention mechanisms. Wang *et al.* (Wang et al. 2023) adopt the contrastive learning strategy to encourage the interaction of multi-modal features. However, these methods overlook the potential over-dependency of the model on specific modalities, and the changes in the informative validity of each modality across various cases. To this end, our proposed multi-modal adaptive learning module could eliminate such limitations to some extent.

## 2.5D Medical Image Segmentation

Advances in medical imaging have led to widespread clinical usage of 3D medical images like MRI and Computed Tomography (CT). When handling volumetric inputs, two prominent strategies emerge: the first involves dividing the 3D volume into 2D slices, and training 2D CNNs for segmentation; the second expands network architecture with 3D convolutions, enabling segmentation across the entire volumetric dataset. Both methods possess distinct advantages and drawbacks (Zhang et al. 2020b). 2D CNNs reduce computational load and offer faster inference, but can miss the information among adjacent slices. Conversely, 3D CNNs capture volumetric spatial relationships, yet their high computational demands and limited performance on anisotropic MR images limit their practicality.

To bridge the 2D-3D CNN gap, innovative 2.5D segmentation methods are proposed for enhancing volumetric medical image segmentation, via novel architectures or strategies fusing volumetric information into 2D CNNs. A multi-view fusion strategy has been utilized to encompass features from sagittal, coronal, and axial images (Ding et al. 2021; Liu et al. 2022). However, the effectiveness of this strategy is still curtailed when dealing with anisotropic volumes, as it struggles to effectively capture spatial relationships and coherent structural information across different image orientations. Another strategy is to integrate volumetric information by incorporating adjacent slices as multi-channel input (Duan et al. 2019; Zhou et al. 2021; Li et al. 2021b). However, such operation frequently introduces redundant and potentially conflicting data from adjacent slices, leading to potential negative effects on segmentation accuracy, and an increased overfitting risk due to the expanded input dimensionality. Consequently, the current researches focus on inter-slice context extraction, treating 2D slices in a 3D volume as a time series sequence to distill information using recurrent neural networks or attention mechanisms.

# Methods

## Approach Overview

As shown in Fig. 3, the proposed model is designed based on the encoder-decoder structure, which mainly consists of two modules: 1) modality-informativeness adaptive learning (MIAL) module, and 2) activation mapping guidance (AMG) module. Given $N$ consecutive slices with three modalities (including T2WI, DWI, and ADC),a two-layer convolution operation is performed to obtain the representation of each modality. As T2WI, DWI, and ADC are distinctly represented and have their own emphases, the initial encoding stage uses a separate encoding branch for each modality. The encoded representations of the three modalities are then fed into the MIAL, which judiciously integrates informative components to mitigate the impact of noisy features and modalities. Considering tumor(s) often appearing in neighboring slices, we borrow the sequence layer (Blattmann et al. 2023) into the deep encoder and decoder to coarsely capture the inter-slice relationship on a global scale, and then we fed the features into AMG module to refine the slice-wise features with low confidence by adopting leverage activation mapping. Notably, with loss of generalizability, we employ the Unet (Hung et al. 2015) architecture as our backbone.

## Modality-Informativeness Adaptive Learning

For multi-modal MR images collected from different subjects, the informativeness of each modality, and each feature extracted from a modality, can vary significantly. Therefore, it is crucial to adaptively enhance informative modalities and features, while suppressing confusing ones, to promote stability in cross-modality and within-modality representation.

**Masked Image Modeling** Let $\mathbf{X}_T \in \mathbb{R}^{H\times W\times C}$, $\mathbf{X}_D \in \mathbb{R}^{H\times W\times C}$, and $\mathbf{X}_A \in \mathbb{R}^{H\times W\times C}$ represent the features extracted by the initial encoder from three modalities such as T2WI, DWI, and ADC, where $H$, $W$, and $C$ denote the height, width, and channel numbers, respectively. In order to comprehensively exploit discriminative features within each modality and mitigate the network's potential modality-dependent bias, we introduce a mask modeling strategy. The features of three modalities are divided equally into multiple

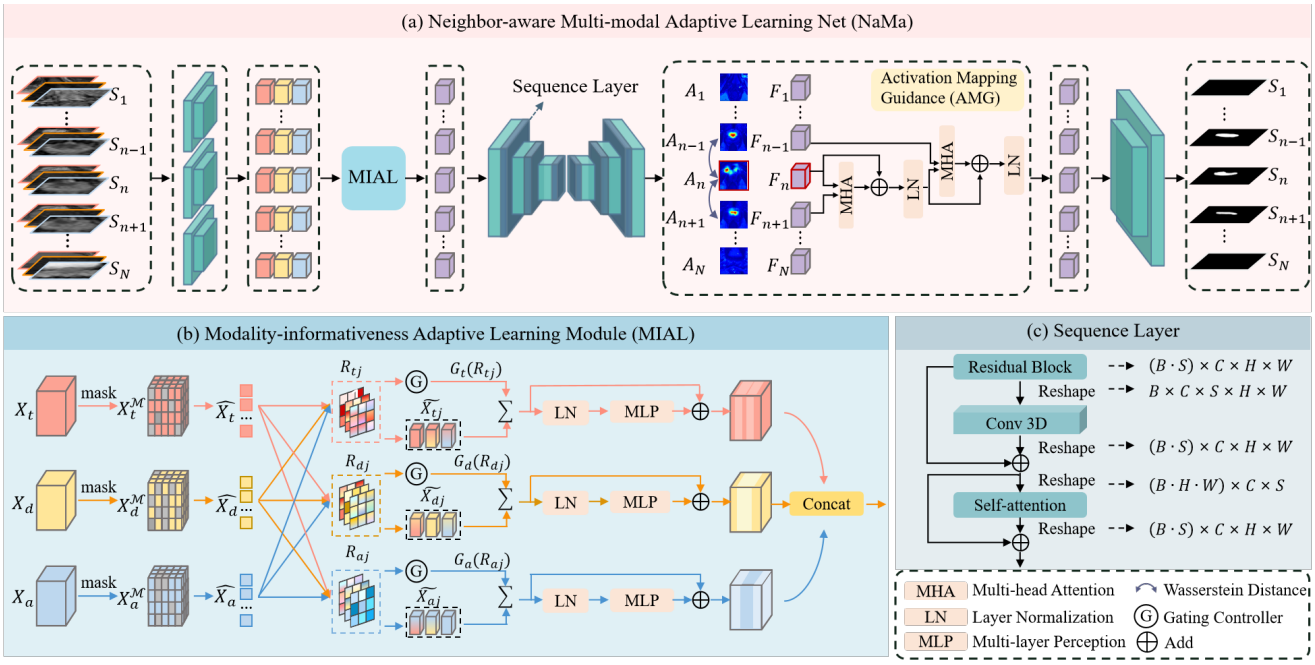


Figure 3: (a) Overview of the proposed NaMa model composed of MIAL and AMG modules. The AMG is adopted to refine the slice-wise low-confidence features using features from adjacent slices. (b) The detailed architecture of MIAL, where the effectiveness of each modality is adaptively evaluated and then cross-modality features are fused accordingly. (c) The detailed structure of sequence layer, which can coarsely capture inter-slice relationship at the global scale.

non-overlapping sub-patches, each of which is subsequently subjected to random masking. The mask $\mathcal{M}_i$ with $i \in \{T, D, A\}$ is drawn from a uniform distribution followed by a predefined thresholding process. The masked $i$-th modality $\mathbf{X}_i^{\mathcal{M}}$ is then obtained by element-wise multiplication between the mask and image by:

$$\mathbf{X}_i^{\mathcal{M}} = \mathbf{X}_i \odot \mathcal{M}_i. \tag{1}$$

The network has to learn to predict the masked regions by considering contextual relationship within each modality. In such a way, our model obtains improved representation learning. Furthermore, by masking random regions in each modality, this procedure contributes to alleviating the influence of redundant information and perplexing elements, consequently enhancing the overall performance.

**Modality-wise Adaptive Learning** After feature embedding on the encoded representations $\mathbf{X}_i^{\mathcal{M}}$, we can obtain the embedded modalities $\widehat{\mathbf{X}}_i$, and then the inter-modality relevance $\mathbf{R}_{ij}$ between $\widehat{\mathbf{X}}_i$ and $\widehat{\mathbf{X}}_j$ can be calculated using a learnable matrix $I$ by:

$$\mathbf{R}_{ij} = \widehat{\mathbf{X}}_i \mathbf{I} \widehat{\mathbf{X}}_j^{\top}, \quad \mathbf{I} \in \mathbb{R}^{C \times C}. \tag{2}$$

High correlation coefficients in $\mathbf{R}_{ij}$ denote high relevance between the corresponding representations in different modalities, indicating the presence of informative modalities. Conversely, low correlation coefficients suggest the presence of at least one modality with confusing regions. Then, $\mathbf{R}_{ij}$ is used as the correlation map to refine each modality. Formally, the correlation-weighted modality $\widetilde{\mathbf{X}}_{ij}$ can be obtained by:

$$\widetilde{\mathbf{X}}_{ij} = \widehat{\mathbf{X}}_j \mathbf{R}_{ij}. \tag{3}$$

To mitigate the adverse impact of uninformative modalities on the informative ones, we develop a gating controller $G_i(\cdot)$ to selectively inhibit uninformative modalities from being transferred into the informative modality branch. The gate layer comprises fully connected layers that utilize soft-max with a small temperature (Maddison, Mnih, and Teh 2016) as an activation function $\sigma$. For uninformative modality $\mathbf{X}_i$, its representations are transferred on its own branch in the following network, due to its low correlation coefficient with other modalities. In contrast, for informative modalities, we fuse informative features across their branches with the weights provided by the gating controller. The fused modality representations $\widetilde{\mathbf{X}}_i$ can be obtained by:

$$\widetilde{\mathbf{X}}_i = \sum_{j \in \{T,D,A\}} \widetilde{\mathbf{X}}_{ij} \odot \sigma\left[G_i(\mathbf{R}_{ij})\right]. \tag{4}$$

Finally, we can obtain the fused informative modality features by performing multi-layer perceptron (MLP) and residual operation:

$$\widetilde{\mathbf{X}}_i' = \widetilde{\mathbf{X}}_i + \mathrm{MLP}(\widetilde{\mathbf{X}}_i). \tag{5}$$

## Activation Mapping Guidance Module

To further ensure the consistency of volumetric segmentation, we evaluate the feature confidence to identify features in need of refinement. According to the confidence

scores, we input inadequately qualified features, alongside their neighbors, into the network to generate a refined slice-wise representation. Given that activation maps depict the network's focus on target regions, we leverage these activation maps to conduct an assessment of feature confidence.

To be specific, as shown in Fig. 3(a), we can obtain features $\mathbf{F}_n \in \{\mathbf{F}_1, ..., \mathbf{F}_l\}$ from consecutive slices within a subject, along with their corresponding activation maps $\mathbf{A}_n \in \{\mathbf{A}_1, ..., \mathbf{A}_l\}$. For the feature $\mathbf{F}_n$, its confidence score can be obtained through the calculation of the Wasserstein distances $\mathcal{WD}_1$ between $\mathbf{F}_n$ and $\mathbf{F}_{n-1}$ *as well as* $\mathcal{WD}_2$ between $\mathbf{F}_n$ and $\mathbf{F}_{n+1}$. Given the inherent continuity of tumors, it is highly likely that the tumor regions across neighboring slices exhibit considerable similarity, with small Wasserstein distances. Therefore, when any $\mathcal{WD}$ surpasses the threshold, it signifies low confidence in the reliability of $\mathbf{F}_n$. It should be noted that activation maps only provide a confidence evaluation of slice-wise features with no gradient back propagation.

Assuming $\mathbf{F}_n$ is evaluated as the features with low confidence, then we can utilize the closest slice-wise features $\mathbf{F}_{n-1}$ and $\mathbf{F}_{n+1}$ as teachers to refine it. To make the network smoothly refine the target feature, we treat the feature with smaller $\mathcal{WD}$ as $\mathcal{T}_1$ and the other is $\mathcal{T}_2$. Subsequently, $\mathcal{T}_1$ is used as key and value, and $\mathbf{F}_n$ is used as query for the attention scheme:

$$\mathbf{Q}_n = \mathbf{F}_n\mathbf{W}_Q, \mathbf{K} = \mathcal{T}_1\mathbf{W}_K, \mathbf{V} = \mathcal{T}_1\mathbf{W}_V, \tag{6}$$

where $\mathbf{W}_Q$, $\mathbf{W}_K$, and $\mathbf{W}_V$ are weights of different inputs of attention scheme. Then, the initial guidance information $\mathbf{GI}_n$ and refined feature are calculated by:

$$\mathbf{GI}_n = \sigma\left[\psi\left(\frac{{\mathbf{Q}_n}^\top\mathbf{K}}{\sqrt{C}}\right)\right]\mathbf{V}^\top, \tag{7}$$

$$\widetilde{\mathbf{F}_n} = \mathbf{GI}_n + \mathrm{MLP}(\mathbf{Q}_n + \mathbf{GI}_n)\,. \tag{8}$$

After that, the initial refined feature $\widetilde{\mathbf{F}_n}$ is used as query, and $\mathcal{T}_2$ is used as key and value for the further refinement by utilizing the same strategy as $\mathcal{T}_1$ does.

## Loss Function

Following previous methods (Zhang et al. 2022a; Zhuang et al. 2022; Yang et al. 2023), we formulate our training loss ($\mathcal{L}$) by combining the Dice loss ($\mathcal{L}_{Dice}$) and the Focal loss ($\mathcal{L}_F$) with deep supervision. Given predicted mask $\hat{Y}$, ground-truth mask $Y$, and voxel number $P$, the loss function is formulated as follows.

$$\begin{aligned}\mathcal{L} &= \mathcal{L}_{Dice} + \mathcal{L}_F \\ &= 1 - \frac{2\sum_{p=1}^{P} y_p\hat{y}_p}{\sum_{p=1}^{P}(y_p + \hat{y}_p + \epsilon)} \\ &\quad - \sum_{p=1}^{P}(\lambda y_p \log\hat{y}_p + (1-\lambda)(1-y_p)\log(1-\hat{y}_p)),\end{aligned} \tag{9}$$

where $y_p \in Y$ and $\hat{y}_p \in \hat{Y}$ are their $p$-th pixel. $\epsilon$ is a smoothing factor for numerical stability, and the weight $\lambda$ is set to 0.95 in our experiments.

# Experiments

## Datasets

Both in-house and public datasets are used to evaluate our proposed model. Specifically, the in-house dataset (IH) comprises multi-modal prostate tumor MR images, encompassing T2WI, DWI, and ADC modalities, sourced from 355 patients. The tumor annotations were meticulously annotated on the axial T2WIs by three proficient radiologists, leveraging insights from biopsy pathology. These annotations underwent thorough cross-validation among the radiologists to establish a definitive and reliable ground truth. Within the IH dataset, we partitioned 249 cases for training, 35 cases for validation, and 71 cases for testing purposes. The public PICAI dataset consists of prostate tumor MR images captured via T2WI, DWI, and ADC modalities, sourced from 215 patients. For the PICAI dataset, we performed a random split, assigning 150 cases for training, 21 cases for validation, and 44 cases for testing.

## Implementation Details

**Data preprocessing.** In order to achieve spatial alignment across three modalities, a rigid transformation is implemented to register the DWIs and ADC maps to the T2WIs. All images are interpolated to the resolution of T2WIs, which is $0.3 \times 0.3 \times 3.0\ mm^3$ for the IH dataset and $0.5 \times 0.5 \times 3.0\ mm^3$ for the PICAI dataset, respectively. The images are cropped into $256 \times 256$ pixels for noise reduction and memory conservation. The effectiveness of image cropping has been demonstrated in our pre-experiment. Detailed information of the pre-experiment can be found in the supplementary material.

**Data augmentation.** To balance the number of normal and tumor slices in the training dataset, we apply various data augmentation techniques (e.g., zoom, horizontal/vertical shift, and gaussian blur) to the tumor slices. Given that 3D methods utilize 3D volumes as input, the consideration of the normal-to-tumor ratio on a slice-by-slice basis becomes unnecessary. Consequently, 3D-wise data augmentation is employed to enhance 3D approaches.

**Two-stage network training.** We first train the network backbone with MIAL module until stable segmentation results are obtained. Subsequently, the sequence layers and the AMG module are added to the segmentation network, and their parameters are optimized in the second training session, where the backbone and MIAL module are fixed. All networks and experiments are implemented using Pytorch on 2 NVIDIA Tesla V100S (40GB) GPUs. All models are trained with mini-batches for 500 epochs using Adam optimizer with an initial learning rate of $5e^{-5}$. The temperature in soft-max activation function is set as 0.1.

**Quantitative analysis.** The segmentation maps are quantitatively evaluated using four metrics, namely the Dice Similariy Coefficient (DSC), 95% Hausdorff Distance (HD95), Average Surface Distance (ASD), and mean Intersection over Union (mIoU). To make the results more convincing, we conduct each experiment for five times and report the mean and standard deviation of each metric.

| Method | IH | | | | PICAI | | | |
|---|---|---|---|---|---|---|---|---|
| | DSC [%]↑ | HD95 [mm]↓ | ASD [mm]↓ | mIoU [%]↑ | DSC [%]↑ | HD95 [mm]↓ | ASD [mm]↓ | mIoU [%]↑ |
| 2D nnUnet | 54.28 ± 0.79 | 15.93 ± 1.12 | 6.92 ± 0.45 | 42.43 ± 0.56 | 47.15 ± 0.38 | 21.67 ± 2.19 | 9.52 ± 0.78 | 34.59 ± 0.48 |
| 3D nnUnet | 58.48 ± 0.49 | 10.26 ± 1.41 | 4.75 ± 0.26 | 46.46 ± 0.28 | 55.24 ± 0.24 | 14.60 ± 1.46 | 5.93 ± 0.32 | 42.14 ± 0.16 |
| ProCDet | 54.48 ± 0.48 | 15.26 ± 1.68 | 7.75 ± 0.68 | 42.46 ± 0.32 | 46.19 ± 0.44 | 23.47 ± 2.26 | 13.15 ± 0.44 | 34.43 ± 0.26 |
| MFSL-Net | 58.56 ± 0.23 | 13.00 ± 0.89 | 4.19 ± 0.41 | 45.37 ± 0.39 | 49.02 ± 0.30 | 14.01 ± 0.98 | 6.02 ± 0.24 | 36.45 ± 0.38 |
| CSAD | 59.06 ± 0.29 | 12.04 ± 0.46 | 3.69 ± 0.30 | 46.07 ± 0.16 | 55.89 ± 0.25 | 13.12 ± 0.82 | 5.01 ± 0.34 | 42.77 ± 0.16 |
| F2Net | 58.56 ± 0.22 | 13.00 ± 0.71 | 4.19 ± 0.29 | 45.37 ± 0.18 | 54.42 ± 0.22 | 13.45 ± 1.23 | 5.46 ± 0.28 | 41.25 ± 0.20 |
| ACMINet | 57.49 ± 0.32 | 13.24 ± 1.23 | 5.02 ± 0.46 | 44.84 ± 0.28 | 50.17 ± 0.46 | 16.93 ± 1.79 | 7.94 ± 0.45 | 37.94 ± 0.28 |
| CAT-Net | 58.67 ± 0.47 | 12.05 ± 0.82 | 3.31 ± 0.82 | 45.86 ± 0.58 | 52.67 ± 0.27 | 16.45 ± 1.07 | 7.03 ± 0.26 | 40.98 ± 0.36 |
| **Ours** | **63.76** ± **0.24** | **8.64** ± **0.44** | **2.13** ± **0.21** | **49.46** ± **0.14** | **59.01** ± **0.17** | **11.95** ± **0.79** | **3.99** ± **0.22** | **45.27** ± **0.19** |

Table 1: Results of comparison experiments, where the best results are in bold.

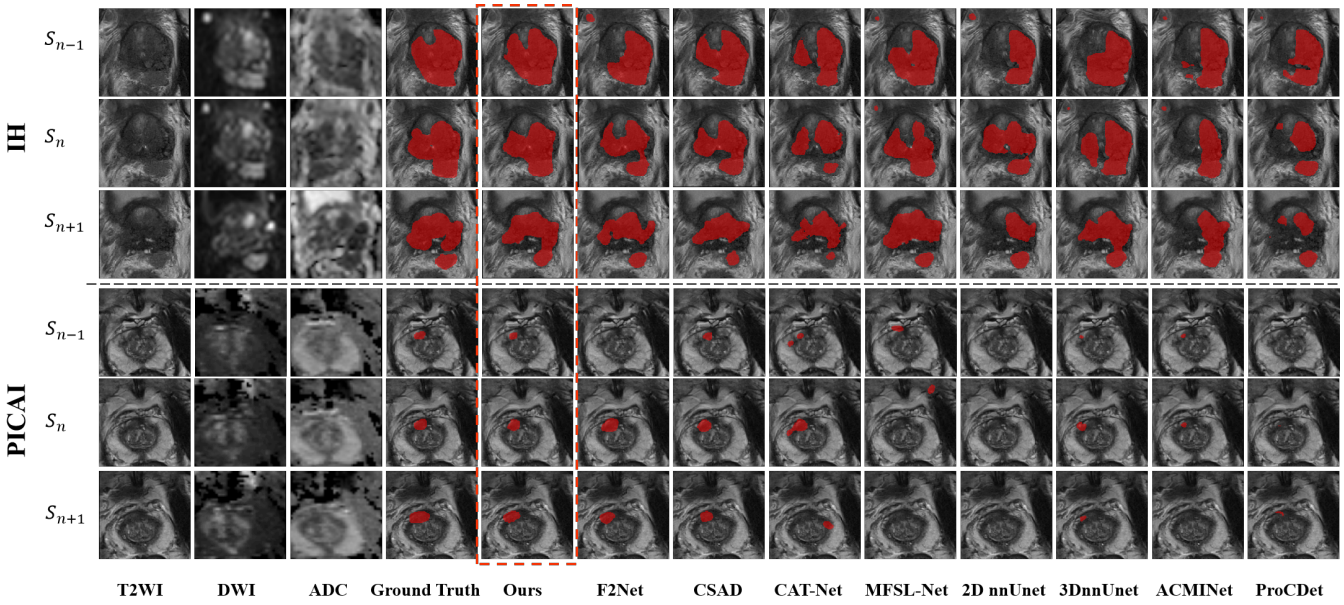


Figure 4: Comparison of tumor segmentation results by different methods on two representative cases in three consecutive slices. The segmentation results are overlaid on T2WI images.

## Comparison with State-of-the-art Methods

To validate the effectiveness of our proposed model, we conduct comparative experiments with several state-of-the-art segmentation methods on both IH and PICAI datasets, and present quantitative and qualitative comparison results in Table 1 and Fig. 4, respectively. These methods can be roughly divided into three categories: 1) universal segmentation methods, including 2D nnUnet, and 3D nnUnet (Isensee et al. 2021); 2) previous prostate tumor segmentation approaches, including ProCDet (Qian, Zhang, and Wang 2021), MFSL-Net (Zhang et al. 2021) and CSAD (Zhang et al. 2022a); and 3) other multi-modal and cross-slice interaction techniques, including F2Net (Yang et al. 2023), ACMINet (Zhuang et al. 2022), and CAT-Net (Hung et al. 2022). For methods not designed for multi-modalities, we adapt them by concatenating three modalities as multi-channel inputs while maintaining the original network structure.

The quantitative results are provided in Table. 1 with the best results boldfaced. Our model achieves the best performance in terms of all evaluated metrics, even against the state-of-the-art CSAD (63.76% vs. 59.06% in DSC on IH and 59.01% vs. 55.89% in DSC on PICAI), demonstrating its advantage in multi-modal MRI prostate tumor segmentation. It is worth noting that our approach showcases substantial superiority over other multi-modal segmentation techniques with cross-slice interaction. For example, in comparison with ACMINet, we achieved noteworthy enhancements in DSC values of 6.27% and 8.84% on the IH and PICAI datasets, respectively. These improvements can be attributed to the elimination of modality-independent bias and the adept capture of the global-to-local cross-slice relationship in anisotropic MR slices.

We further present qualitative results of two typical cases in Fig. 4, each consisting of three consecutive slices. In the case with sizable tumor (the first case), all methods success-

| Method | IH | | | | PICAI | | | |
|---|---|---|---|---|---|---|---|---|
| | DSC [%]↑ | HD95 [mm]↓ | ASD [mm]↓ | mIoU [%]↑ | DSC [%]↑ | HD95 [mm]↓ | ASD [mm]↓ | mIoU [%]↑ |
| bNet-T | 39.28 ± 0.34 | 20.28 ± 0.89 | 8.43 ± 0.25 | 20.64 ± 0.34 | 32.41 ± 0.32 | 37.30 ± 0.84 | 18.95 ± 0.32 | 14.66 ± 0.31 |
| bNet-D | 43.32 ± 0.28 | 21.81 ± 0.54 | 8.47 ± 0.25 | 23.53 ± 0.38 | 38.78 ± 0.34 | 29.42 ± 0.68 | 12.02 ± 0.29 | 26.98 ± 0.29 |
| bNet-A | 43.73 ± 0.29 | 19.60 ± 0.59 | 8.38 ± 0.34 | 24.61 ± 0.35 | 38.32 ± 0.45 | 29.30 ± 0.76 | 13.66 ± 0.33 | 27.31 ± 0.39 |
| bNet-Full | 56.44 ± 0.16 | 12.39 ± 0.46 | 4.26 ± 0.13 | 43.14 ± 0.14 | 50.55 ± 0.13 | 18.59 ± 0.32 | 7.47 ± 0.19 | 38.58 ± 0.11 |
| bNet-Full-M | 61.57 ± 0.15 | 11.13 ± 0.41 | 3.32 ± 0.12 | 47.02 ± 0.16 | 57.44 ± 0.09 | 15.20 ± 0.38 | 5.62 ± 0.11 | 43.09 ± 0.12 |
| bNet-Full-A | 60.63 ± 0.11 | 10.45 ± 0.24 | 2.95 ± 0.18 | 46.30 ± 0.17 | 55.16 ± 0.14 | 13.80 ± 0.32 | 4.93 ± 0.17 | 41.79 ± 0.14 |
| bNet-Full-M-A | **63.76 ± 0.24** | **8.64 ± 0.44** | **2.13 ± 0.21** | **49.46 ± 0.14** | **59.01 ± 0.17** | **11.95 ± 0.79** | **3.99 ± 0.22** | **45.27 ± 0.19** |

Table 2: Ablation study of key components. The best results are in bold.

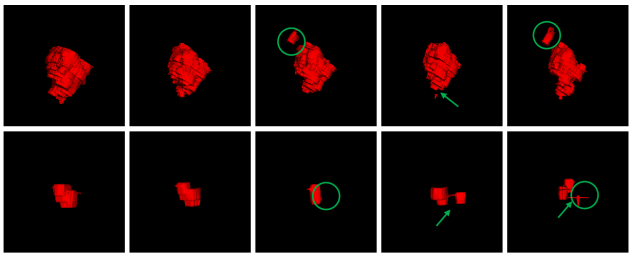


Figure 5: 3D visualization for the results of our proposed method using different components. The two rows show results from the same two cases as Fig. 4. The part circled in green represents the over- or under-segmented regions and arrows point to discontinuous segmentations.

fully delineate a portion of the tumor. But our model is more consistent with the ground truth, *not only* in the global shape *but also* in the local details. For the challenging case (the second case), our proposed method outperforms the competing methods by a significant margin. Intuitively, the possible reason is that MIAL effectively explores informative modalities and features, and AMG efficiently enhances inter-slice segmentation consistency.

## Ablation Studies

We perform ablation studies to evaluate the effective of each component in the whole segmentation framework. This involves utilizing three different configurations of a baseline Unet model with sequence layers for tumor segmentation: 1) from a single modality (bNet-T, bNet-D, bNet-A for T2WI, DWI and ADC map, respectively), 2) from three concatenated modalities (bNet-Full), and 3) from three modalities after adding MIAL module (bNet-Full-M) or the AMG module (bNet-Full-A) .

As shown in Table 2, the multi-modal information can significantly improve segmentation accuracy (DSC) from 39.28% to 63.76% on IH and 32.41% to 59.01% on PICAI, compared to the case of using single-modality. Moreover, additional components have a significant impact on improving performance of the model in terms of segmentation results. As expected, the incorporation of the MIAL aids the network in prioritizing informative aspects of different modalities and features, resulting in improved performance in accurately capturing the shape and location of prostate tumors in multi-modal MR images, as demonstrated by higher DSC scores (5.13% improvement on IH and 6.89% improvement on PICAI) and higher mIoU scores (3.88% improvement on IH and 4.51% improvement on PICAI). This observation can also be confirmed by the visualization result from the fourth column of Fig. 5. Additionally, incorporating AMG module leads to outstanding improved performance in HD95 (10.45 mm vs. 12.39 mm on IH and 13.80 mm vs. 18.59 mm on PICAI) and ASD (2.95 mm vs. 4.26 mm on IH and 4.93 mm vs. 7.47 mm on PICAI), demonstrating its ability in capturing the cross-slice relationship and aiding the model in generating consistent results. The incorporation of MIAL with AMG yields substantial improvements for all evaluation metrics compared to the baseline.

As shown in Fig. 5, we present the 3D visualization results for two subjects, showcasing the segmentation performance of the ablation study. The presence of green circles and arrows indicate those inaccurately segmented tumors, highlighting the subpar performance of the baseline model without considering the multi-modal adaptive fusion and cross-slice relationship. In contrast, with incorporation of the MIAL and AMG modules, a noticeable enhancement in contour and shape consistency can be observed. These findings underscore significance of capturing complementary information across multiple modalities and cross-slice relationship in prostate tumor segmentation.

## Conclusion

In this paper, we have proposed a novel model, namely neighbor-aware multi-modal adaptive learning network, for automatic prostate tumor segmentation in multi-modal anisotropic MR images. Our model adaptively selects effective modalities and features for each subject by the proposed MIAL module. Furthermore, we introduce a novel AMG module to systematically learn and leverage inter-slice information to mitigate segmentation discontinuities across slices in prostate tumors. We have conducted extensive experiments on both in-house and public datasets to evaluate our proposed model quantitatively and qualitatively from various perspectives. It is shown that our method outperforms the existing state-of-the-art segmentation methods by a large margin. Our future study will focus on investigating the impact of prostate tumor segmentation performance on surgical treatment, radiotherapy planning, and predictive analysis.

## Acknowledgements

This work was supported in part by National Natural Science Foundation of China (grant numbers 62131015, 62250710165), and The Key R&D Program of Guangdong Province, China (grant number 2021B0101420006).

## References

Blattmann, A.; Rombach, R.; Ling, H.; Dockhorn, T.; Kim, S. W.; Fidler, S.; and Kreis, K. 2023. Align your Latents: High-Resolution Video Synthesis with Latent Diffusion Models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Boettcher, A. N.; Usman, A.; Morgans, A.; VanderWeele, D. J.; Sosman, J.; and Wu, J. D. 2019. Past, current, and future of immunotherapies for prostate cancer. *Frontiers in oncology*, 9: 884.

Chen, J.; Wan, Z.; Zhang, J.; Li, W.; Chen, Y.; Li, Y.; and Duan, Y. 2020. Medical image segmentation and reconstruction of prostate tumor based on 3D AlexNet. *Computer methods and programs in biomedicine*, 105878.

Cornud, F.; Delongchamps, N. B.; Mozer, P.; Beuvon, F.; Schull, A.; Muradyan, N.; and Peyromaure, M. 2012. Value of multiparametric MRI in the work-up of prostate cancer. *Current urology reports*, 13: 82–92.

Ding, Y.; Zheng, W.; Geng, J.; Qin, Z.; Choo, K.-K. R.; Qin, Z.-Q.; and Hou, X. 2021. MVFusFra: A Multi-View Dynamic Fusion Framework for Multimodal Brain Tumor Segmentation. *IEEE Journal of Biomedical and Health Informatics*, 26: 1570–1581.

Duan, J.; Bello, G.; Schlemper, J.; Bai, W.; Dawes, T. J. W.; Biffi, C.; de Marvao, A.; Doumoud, G.; O'Regan, D. P.; and Rueckert, D. 2019. Automatic 3D Bi-Ventricular Segmentation of Cardiac Images by a Shape-Refined Multi- Task Deep Learning Approach. *IEEE Transactions on Medical Imaging*, 38(9): 2151–2164.

Havaei, M.; Davy, A.; Warde-Farley, D.; Biard, A.; Courville, A. C.; Bengio, Y.; Pal, C. J.; Jodoin, P.-M.; and Larochelle, H. 2015. Brain tumor segmentation with Deep Neural Networks. *Medical Image Analysis*, 35: 18–31.

Hung, A. L. Y.; Zheng, H.; Miao, Q.; Raman, S. S.; Terzopoulos, D.; and Sung, K. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. *Lecture Notes in Computer Science*, 234-241.

Hung, A. L. Y.; Zheng, H.; Miao, Q.; Raman, S. S.; Terzopoulos, D.; and Sung, K. 2022. CAT-Net: A Cross-Slice Attention Transformer Model for Prostate Zonal Segmentation in MRI. *IEEE Transactions on Medical Imaging*, 42: 291–303.

Isensee, F.; Jaeger, P. F.; Kohl, S. A.; Petersen, J.; and Maier-Hein, K. H. 2021. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2): 203–211.

Jiale, L.; Hang, D.; Hao, H.; and Yong, D. 2023. MSeg3D: Multi-Modal 3D Semantic Segmentation for Autonomous Driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 21694–21704.

Li, J.; Cui, Z.; Wang, S.; Wei, J.; Feng, J.; Liao, S.; and Shen, D. 2021a. Morphology-Guided Prostate MRI Segmentation with Multi-slice Association. In *MLMI@MICCAI*.

Li, L.; Lian, S.; Luo, Z.; Li, S.; Wang, B.; and Li, S. 2021b. Learning Consistency- and Discrepancy-Context for 2D Organ Segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*.

Liu, D.; Gao, Y.; Zhangli, Q.; Wen, S.; Yan, Z.; Zhou, M.; and Metaxas, D. N. 2022. TransFusion: Multi-view Divergent Fusion for Medical Image Segmentation with Transformers. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*.

Maddison, C. J.; Mnih, A.; and Teh, Y. W. 2016. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*.

Man, Y.; Gui, L.-Y.; and Wang, Y.-X. 2023. BEV-Guided Multi-Modality Fusion for Driving Perception. In *CVPR*.

Pereira, S.; Pinto, A.; Alves, V.; and Silva, C. A. 2016. Brain Tumor Segmentation Using Convolutional Neural Networks in MRI Images. *IEEE Transactions on Medical Imaging*, 35: 1240–1251.

Qian, Y.; Zhang, Z.; and Wang, B. 2021. ProCDet: A New Method for Prostate Cancer Detection Based on MR Images. *IEEE Access*, 9: 143495–143505.

Tanimoto, A.; Nakashima, J.; Kohno, H.; Shinmoto, H.; and Kuribayashi, S. 2007. Prostate cancer screening: the clinical value of diffusion-weighted imaging and dynamic MR imaging in combination with T2-weighted imaging. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 25(1): 146–152.

Velonas, V. M.; Woo, H. H.; dos Remedios, C. G.; and Assinder, S. J. 2013. Current status of biomarkers for prostate cancer. *International journal of molecular sciences*, 14(6): 11034–11060.

Wang, Y.; Peng, J.; Zhang, J.; Yi, R.; Wang, Y.; and Wang, C. 2023. Multimodal Industrial Anomaly Detection via Hybrid Fusion.

Xing, Z.-Y.; Yu, L.; Wan, L.; Han, T.; and Zhu, L. 2022. NestedFormer: Nested Modality-Aware Transformer for Brain Tumor Segmentation. *ArXiv*, abs/2208.14876.

Yang, H.; Zhou, T.; Zhou, Y.; Zhang, Y.; and Fu, H. 2023. Flexible Fusion Network for Multi-Modal Brain Tumor Segmentation. *IEEE Journal of Biomedical and Health Informatics*, 27: 3349–3359.

Zhan, B.; Li, D.; Wu, X.; Zhou, J.; and Wang, Y. 2021. Multi-Modal MRI Image Synthesis via GAN With Multi-Scale Gate Mergence. *IEEE Journal of Biomedical and Health Informatics*, 26: 17–26.

Zhang, F.; Zhang, B.; Zhang, Z.; Mi, Y.; Wu, J.; Huang, H.; Que, X.; and Wang, W. 2021. MFSL-Net: A Modality Fusion and Shape Learning based Cascaded Network for Prostate Tumor Segmentation. *2021 IEEE International Conference on Big Data (Big Data)*, 3943–3949.

Zhang, G.; Shen, X.; Zhang, Y.-D.; Luo, Y.; Luo, J.; Zhu, D.; Yang, H.; Wang, W.; Zhao, B.; and Lu, J. 2022a. Cross-Modal Prostate Cancer Segmentation via Self-Attention Distillation. *IEEE Journal of Biomedical and Health Informatics*, 26(11): 5298–5309.

Zhang, Y.; He, N.; Yang, J.; Li, Y.; Wei, D.; Huang, Y.; Zhang, Y.; He, Z.; and Zheng, Y. 2022b. mmFormer: Multimodal Medical Transformer for Incomplete Multimodal Learning of Brain Tumor Segmentation. *ArXiv*, abs/2206.02425.

Zhang, Y.; Liao, Q.; Ding, L.; and Zhang, J. 2020a. Bridging 2D and 3D segmentation networks for computation-efficient volumetric medical image segmentation: An empirical study of 2.5D solutions. *Computerized medical imaging and graphics : the official journal of the Computerized Medical Imaging Society*, 99: 102088.

Zhang, Y.; Liao, Q.; Ding, L.; and Zhang, J. 2020b. Bridging 2D and 3D segmentation networks for computation-efficient volumetric medical image segmentation: An empirical study of 2.5D solutions. *Computerized medical imaging and graphics : the official journal of the Computerized Medical Imaging Society*, 99.

Zhang, Y.; Sidibé, D.; Morel, O.; and Mériaudeau, F. 2020c. Deep multimodal fusion for semantic image segmentation: A survey. *Image Vis. Comput.*, 105: 104042.

Zhou, H.; Xiao, J.; Fan, Z.; and Ruan, D. 2021. Intracranial Vessel Wall Segmentation For Atherosclerotic Plaque Quantification. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, 1416–1419.

Zhuang, Y.; Liu, H.; Song, E.; and Hung, C.-C. 2022. A 3D Cross-Modality Feature Interaction Network With Volumetric Feature Alignment for Brain Tumor and Tissue Segmentation. *IEEE Journal of Biomedical and Health Informatics*, 27: 75–86.