Volumetric Axial Disentanglement Enabling Advancing in Medical Image Segmentation

Xingru Huang[1], **Jian Huang**[1], **Yihao Guo**[1], **Tianyun Zhang**[1], **Zhao Huang**[2], **Yaqi Wang**[3], **Ruipu Tang**[4], **Guangliang Cheng**[5], **Shaowei Jiang**[1*], **Zhiwen Zheng**[1*], **Jin Liu**[1*], **Renjie Ruan**[6*], **Xiaoshuai Zhang**[7*]
[1]Hangzhou Dianzi University, Hangzhou, China
[2]Northumbria University, Newcastle, UK
[3]Communication University of Zhejiang, Hangzhou, China
[4]Beijing University of Posts and Telecommunications, Beijing, China
[5]University of Liverpool, Liverpool, UK
[6]The Third Affiliated Hospital of Wenzhou Medical University, Wenzhou, China
[7]Ocean University of China, Qingdao, China
{xingru.huang, j.huang, yihao.guo, tianyun.zhang, jiangsw, zhiwen.zheng, jinliu}@hdu.edu.cn,
zhao.huang@northumbria.ac.uk, wangyaqi@cuz.edu.cn, tangruipu0817@bupt.edu.cn,
guangliang.cheng@liverpool.ac.uk, ruanrenjie@wmu.edu.cn, x.zhang@ouc.edu.cn

Abstract

Information retrieved from three dimensions is treated uniformly in CNN-based volumetric segmentation methods. However, such neglect of axial disparities fails to capture true spatio-temporal variations. This paper introduces the volumetric axial disentanglement to address the disparities in spatial information along different axial dimensions. Building on this concept, we propose the Post-Axial Refiner (PaR) module to refine segmentation masks by implementing axial disentanglement on the specific axis of the volumetric medical sequences. As a plug-and-play enhancement to existing volumetric segmentation architecture, PaR further utilizes specialized attention approaches to learn disentangled post-decoding features, enhancing spatial representation and structural detail. Validation on various datasets demonstrates PaR's consistent elevation of segmentation precision and boundary clarity across 11 baselines and different imaging modalities, achieving state-of-the-art performance on multiple datasets. Experimental tests demonstrate the ability of volumetric axial disentanglement to refine the segmentation of volumetric medical images. Code is released at https://github.com/IMOP-lab/PaR-Pytorch.

1 Introduction

The precise delineation of clinical diagnoses critically hinges on the analysis of volumetric medical imaging data, such as MRI and CT scans. These volumetric data encapsulate the exact spatial configurations of tissues and organs within patients' bodies and may capture dynamic changes over time, including blood flow or organ movement. The inherent volumetric nature of these data enables an in-depth analysis of pathological states and physiological processes. Therefore, accurate segmentation of complex tissue structures is paramount for extracting precise biomorphic information, effectively marking disease, and facilitating subsequent analyses and diagnoses. However, 2D segmentation methods fail to address the hierarchical relationships inherent in these data since they lack the capacity to adequately represent the spatial continuity of volumetric data to incur segmentations with poor 3D cohesion and precision.

The symmetrical design of 3D conv extends the convolution kernel across all axes to extract features from the entire volumetric space by treating the time axis (z-axis) information the same as cross-section (xy-plane) information, and hence to optimize the use of continuous volume information. This approach overlooks the distinct physical and scanning characteristics of different axes, especially for the differences between the z-axis and the xy-plane in terms of physical properties and information density. As illustrated in Fig. 1, the visualization of volumetric data highlights the differing information densities across axes, the xy-plane accurately reflects tissue density distribution and shows precise locations of various tissue types, and the z-axis represents the morphological changes of the scanned sequence over time. When z-axis information is rich, the model should focus more on its detailed z-axis features; whilst z-axis information is sparse, the model should prioritize the detailed features of the xy-plane. However, the current 3D conv approaches fail to consider these differences, often leading to inadequate differentiation and ineffective use of the unique characteristics of each axis.

Due to the differences in resolution and information density between the z-axis and xy-plane in volumetric data, the

[*] Corresponding authors.

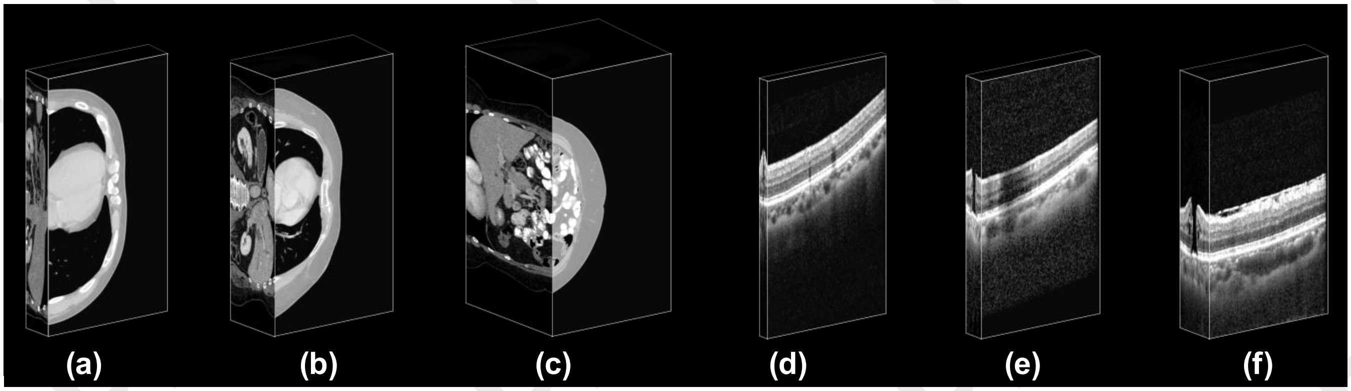

Figure 1: **Visualization of different volumetric data.** Abdomen CT volumes (a-c) and retinal OCT volumes (d-f) show significant variations in different axial lengths relative to actual pixel dimensions, highlighting the differences in information density across axes.

usage of 3D conv may introduce spatial biases and temporal errors by uniformly extracting features across all dimensions. This can lead to the misapplication of xy-plane features to the temporal z-axis, thereby limiting segmentation accuracy. To address this issue, we propose the volumetric axial disentanglement and further design the **Post-Axial Refiner (PaR)** module, which disentangles specific axial features and applies specialized attention approaches to post-decoding 3D feature maps (coarse masks). PaR module refines temporal and spatial features for accurate mapping of temporal variations and structural differences to enhance segmentation accuracy. **This plug-and-play module can be seamlessly integrated into any 3D segmentation network architecture.**

The proposed method is validated using publicly available volumetric segmentation datasets (FLARE2021, OIMHS, and SegTHOR) and compared with 11 previous state-of-the-art models to verify its effectiveness. The results demonstrate effective improvements in segmentation performance across all models with the integration of PaR, including enhanced overall accuracy and boundary delineation. The results underscore the enhancement of volumetric axial disentanglement for precise medical volumetric segmentation.

Our contributions are summarized as follows:

1. Substantiating the operation of volumetric axial disentanglement as a means to tackle the challenges arising from variations in physical properties and information density across distinct axial directions in 3D medical imaging.

2. Introducing a simple, plug-and-play PaR module that performs volumetric axial disentanglement and attention-based axial feature extraction to enhance specific axial feature representation in 3D medical imaging.

3. By validating on multiple publicly available datasets FLARE2021, OIMHS, and SegTHOR with 11 of the mainstream volumetric segmentation models, the proposed PaR module consistently improves segmentation performance.

2 Related Work

2.1 Volumetric Segmentation

Volumetric segmentation methods have significantly advanced due to developments in deep learning and computational resources. The rise and continuous iterations of Convolutional Neural Networks (CNN) and Vision Transformers (ViT) [Dosovitskiy *et al.*, 2020] have propelled these advancements [Liao *et al.*, 2020; Huang *et al.*, 2023; Huang *et al.*, 2024; Huang *et al.*,]. The 3D U-Net [Çiçek *et al.*, 2016], a seminal model in this field, employs 3D conv to capture spatial relationships within volumetric data. Enhancements to the 3D U-Net include attention mechanisms that prioritize relevant regions of the input data [Jiang *et al.*, 2022] to enhance feature extraction capabilities and reduce noise interference. Transformer [Vaswani *et al.*, 2017] models adapted for volumetric segmentation leverage self-attention for long-range dependencies, as seen in the UNETR architecture [Hatamizadeh *et al.*, 2021a; Shaker *et al.*, 2024]. Hybrid models that integrate CNNs and transformers balance spatial feature extraction with global context capture, resulting in further performance gains [Chen *et al.*, 2021; Wang *et al.*, 2021; Xie *et al.*, 2021; Hatamizadeh *et al.*, 2021b]. Notably, the nnU-Net [Isensee *et al.*, 2021], a self-adapting framework, has demonstrated superior performance in a wide range of medical imaging tasks.

Beyond network architecture, numerous other research efforts have emerged for volumetric segmentation. Post-processing approaches, including morphological operations and conditional random fields (CRFs), are employed to refine segmentation outcomes [Chen *et al.*, 2022]. Semi-supervised and unsupervised segmentation methods enhance robustness by leveraging large volumes of unlabeled data [Bortsova *et al.*, 2019; Huang *et al.*, 2022]. Additionally, Generative Adversarial Networks (GANs) improve the realism and accuracy of segmentation outputs [Kwon *et al.*, 2019; Goodfellow *et al.*, 2020]. Federated learning enables model training across institutions while preserving privacy [Kaissis *et al.*, 2021; Miao *et al.*, 2023]. Overall, continuous innovations in neural network architectures and learning paradigms promise more accurate and reliable medical image analysis.

2.2 Segmentation Mask Refinemement

Segmentation mask refinement methods aim to improve the accuracy of masks generated by pre-existing segmentation models. While numerous refinement methods exist for 2D image segmentation, few of them address the refinement of volumetric medical image segmentation masks.

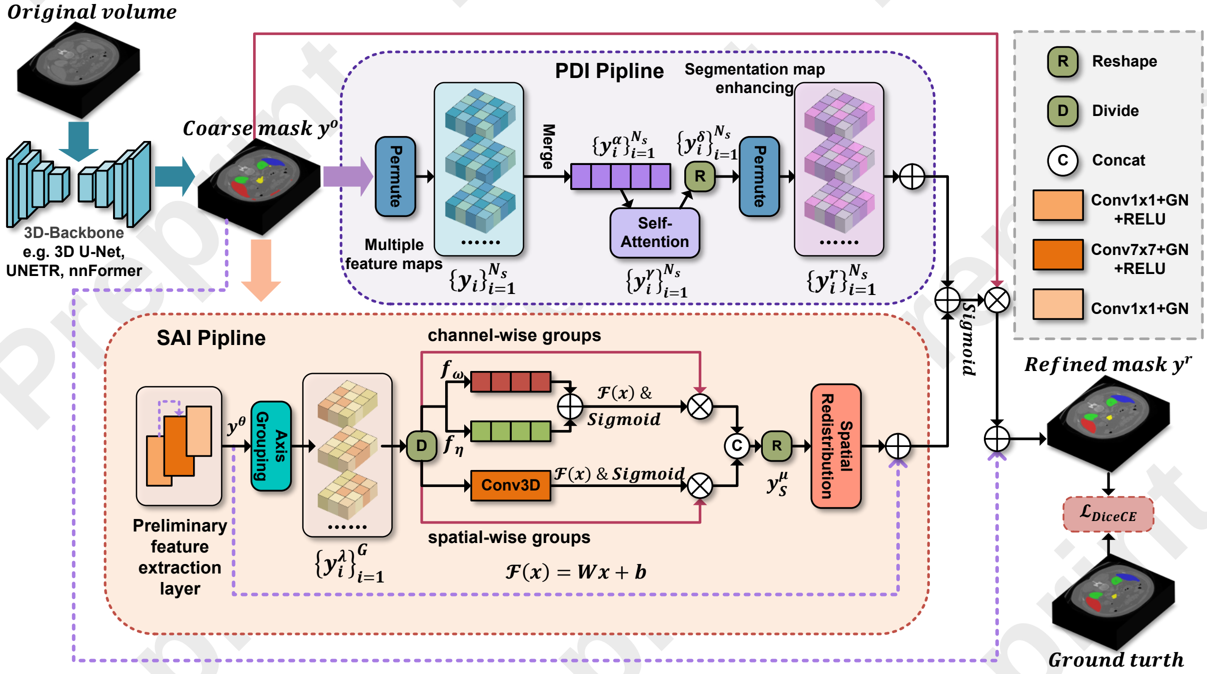

Figure 2: **Overview of the proposed PaR module** (best viewed in color). The PaR module consists of two main pipelines: PDI and SAI. The PDI pipeline applies multi-head self-attention to capture and learn specific axial feature information from various feature map perspectives, while the SAI pipeline focuses on extracting and enhancing specific axial spatio-temporal representation through grouped disentanglement, two attention groups, and spatial redistribution, thereby achieving refinement of the coarse mask.

In 2D segmentation mask refinement, PointRend [Kirillov *et al.*, 2020] employs an MLP to correct low-confidence pixel labels within Mask R-CNN [He *et al.*, 2017]. SegFix [Yuan *et al.*, 2020] improves edge predictions by learning a mapping function between edge and inner pixels to replace inaccurate edge predictions with corresponding inner pixel predictions. RefineMask [Zhang *et al.*, 2021] enhances Mask R-CNN by incorporating a semantic head for additional guidance. Mask-Transfiner [Ke *et al.*, 2022] uses an FCN to detect incoherent regions and refine their labels with a Transformer.

As for the refinement of the volumetric image segmentation mask, [Duan *et al.*, 2019] designs a refinement step that imposes shape prior knowledge to augment segmentation quality and mitigate image artifacts. [Xue *et al.*, 2023] introduces an approach employing a graph convolutional network (GCN) to refine inaccurate segmentation masks. [Shu *et al.*, 2024] proposes a method utilizing a self-refine module to improve the labels produced by an auxiliary network through progressively accurate predictions from the primary network.

3 Methodology

Firstly, we provide a concise overview of the **Post-Axial Refiner (PaR)** module, which consists of **two main pipelines: Permutative Dimensional Intensification (PDI) and Sequential Axial Intensification (SAI)**. We then explore the structure and functionality of the PDI pipeline and analyze its role in capturing and learning information of specific axial features. Finally, we present the SAI pipeline, which is dedicated to axial disentanglement on a specific axis for extracting and enhancing distinctive spatio-temporal characteristics.

3.1 Overview of the Post-Axial Refiner

The detailed structure of PaR is illustrated in Fig. 2. The baseline network takes the original volume image as input, and the 3D feature map generated by its decoder is defined as the coarse mask $y^o \in \mathbb{R}^{C\times W\times H\times D}$, where C represents the number of segmentation categories, and W, H, and D denote the width, height, and depth of the volumetric medical images, respectively. Given the inherently richer spatio-temporal information and texture details of volumetric medical images as opposed to 2D plane images, PaR utilizes volumetric axial disentanglement to extract the key features across the transverse, sagittal, and coronal planes of volumetric data to enhance the model's feature representation capability. To effectively capture the distinct feature variations in each axial plane and extract important feature information, PaR employs two pipelines: PDI and SAI.

The **PDI pipeline** is designed to capture and learn critical feature information within the coarse masks by applying a dimensional permutation operation and multi-head self-attention [Vaswani *et al.*, 2017] along the specific axis. Initially, various spatial distribution tensors are generated from the coarse masks y^o by a permutation operation to enrich the semantic content. Subsequently, multi-head self-attention is employed to learn a unified feature structure across these spatial distributions to ultimately capture the most salient specific axial feature information, represented as $y_P^r \in \mathbb{R}^{C\times W\times H\times D}$.

For the enhancement of specific axial feature representation in volumetric data through volumetric axial disentanglement, we introduce the **SAI pipeline**. It initiates with 3D Conv, Group Normalization (GN), ReLU, and residual connection, preliminarily extracting features from the coarse masks. These features are then grouped along the specific axis and analyzed both channel-wise and spatial-wise to learn the key characteristic of specific axial features, resulting in $y_S^r \in \mathbb{R}^{C\times W\times H\times D}$.

The outputs of both PDI y_P^r and SAI y_S^r are combined through addition, subsequently constrained within the (0,1) range via a sigmoid function, and then element-wise multiplied by the original coarse mask y^o. This product is further enhanced through residual connections with the mathematical formulation of this progress as follows:

$$y^r = \frac{y^o}{1 + e^{-(y_P^r + y_S^r)}} + y^o,$$

where $y^r \in \mathbb{R}^{C\times W\times H\times D}$ is the refined segmentation mask.

3.2 Permutative Dimensional Intensification

Permutative Dimensional Intensification (PDI) improves volumetric image segmentation accuracy and robustness by capturing and learning specific axial feature information across different spatial distributions within the coarse masks. Here, we set the W-axis as the specific disentanglement axis.

For the coarse mask y^o, the PDI pipeline generates spatial distributions through a dimensional permutation operation. This operation rearranges the dimensions of y^o to produce multiple feature maps $\{y_i\}_{i=1}^{N_S}$ under different spatial configurations, such as $y_1 \in \mathbb{R}^{H\times D\times C\times W}$, $y_2 \in \mathbb{R}^{C\times D\times H\times W}$, and $y_3 \in \mathbb{R}^{C\times H\times D\times W}$. The W-axis is permuted to the last dimension to enhance the focus on W-axial information in subsequent attention. These permuted feature maps are then merged into 3D tensors $\{y_i^\alpha\}_{i=1}^{N_S}$ to make them suitable as the input for the multi-head self-attention, which is applied to the extraction of time-axial and spatial features. The formula for multi-head self-attention can be expressed as:

$$\{y_i^\gamma = SelfAttention(y_i^\alpha)\}_{i=1}^{N_S},$$

where $\{y_i^\gamma\}_{i=1}^{N_S}$ are the attention outputs.

Subsequently, attention outputs $\{y_i^\gamma\}_{i=1}^{N_S}$ are reshaped back into 5D formats $\{y_i^\delta\}_{i=1}^{N_S}$ through the reshape operation to be aligned with the dimensions of the coarse masks y^o by permutation, denoted as enhanced segmentation maps $\{y_i^r\}_{i=1}^{N_S}$. The final step in PDI involves a summative integration of these enhanced segmentation maps to produce the W-axial refined segmentation mask y_P^r, thereby strengthening the coarse mask feature representation through capturing and learning specific spatio-temporal information.

3.3 Sequential Axial Intensification

The inherent brevity of sequence lengths in volumetric medical images presents a fundamental drawback that is often mitigated through interpolation in the data preprocessing stage, leading to variations in time-axial information. Given that most segmentation networks typically identify segmentation targets based on the slice plane, they struggle to understand the rich spatio-temporal information across the sequence. To address this issue, we introduce the Sequential Axial Intensification (SAI) pipeline, which applies axial disentanglement and two attention groups along the specific axis (W-axis here). This pipeline enhances W-axial spatio-temporal representation and learns unified features across different axial planes of the coarse masks to refine the coarse masks.

For the coarse mask y^o, the SAI pipeline begins with an initial feature extraction layer comprising 3D Conv, GN, ReLU, and residual connection. This layer performs the preliminary feature extraction to produce the feature map $y^\theta \in \mathbb{R}^{C\times W\times H\times D}$. To capture the unique patterns and characteristics of the W-axial feature by subsequent attention in y^θ, the feature map is divided along the W-axis into subsets $\{y_i^\lambda \in \mathbb{R}^{C\times\frac{W}{G}\times H\times D}\}_{i=1}^G$. These subsets are further divided into two groups: $\{y_i^{\lambda,c} \in \mathbb{R}^{C\times\frac{W}{2G}\times H\times D}\}_{i=1}^G$ and $\{y_i^{\lambda,s} \in \mathbb{R}^{C\times\frac{W}{2G}\times H\times D}\}_{i=1}^G$.

The channel-wise groups. $\{y_i^{\lambda,c}\}_{i=1}^G$ is processed through the adaptive max pooling f_ω and the adaptive mean pooling f_η to extract W-axial channel information, which is then combined via addition and passed through an adaptive weighting function $\mathcal{F}_c(x) = W_c x + b_c$ to generate $\{y_i^{f_c} \in \mathbb{R}^{C\times 1\times 1\times 1}\}_{i=1}^G$:

$$\begin{aligned} y_i^{f_c} &= \mathcal{F}_c(f_\omega(y_i^{\lambda,c}) + f_\eta(y_i^{\lambda,c})) \\ &= W_c(f_\omega(y_i^{\lambda,c}) + f_\eta(y_i^{\lambda,c})) + b_c, \quad \forall i \in \{1,\ldots,G\}, \end{aligned}$$

where $W_c \in \mathbb{R}^{1\times 1\times 1\times 1}$ and $b_c \in \mathbb{R}^{C\times 1\times 1\times 1}$ are trainable parameters. Note that the initial weight of W_c is set to 1, and the initial weight of b_c is set to 0.

Then $\{y_i^{f_c}\}_{i=1}^G$ are constrained between (0,1) using a sigmoid function and combined with $\{y_i^{\lambda,c}\}_{i=1}^G$ through element-wise multiplication to extract the key channel features within $\{y_i^{\lambda,c}\}_{i=1}^G$.

The spatial-wise groups. $\{y_i^{\lambda,s}\}_{i=1}^G$ undergo a 3D Conv layer to capture spatial features within the local neighborhood of the feature map. This operation can reflect the significance of spatial positions across $\{y_i^{\lambda,s}\}_{i=1}^G$. The captured features are then processed by an adaptive weighting function $\mathcal{F}_s(x) = W_s x + b_s$ to generate $\{y_i^{f_s} \in \mathbb{R}^{1\times\frac{W}{2G}\times H\times D}\}_{i=1}^G$:

$$\{y_i^{f_s} = \mathcal{F}_s(Conv3D(y_i^{\lambda,s})) = W_s(Conv3D(y_i^{\lambda,s})) + b_s\}_{i=1}^G,$$

where $W_s \in \mathbb{R}^{1\times 1\times 1\times 1}$ and $b_s \in \mathbb{R}^{1\times\frac{W}{2G}\times H\times D}$ are trainable parameters, set to 1 and 0, respectively.

Similarly, $\{y_i^{f_s}\}_{i=1}^G$ is processed by sigmoid activation to ensure values between (0,1) and then are combined with $\{y_i^{\lambda,s}\}_{i=1}^G$ by element-wise multiplication to extract the key spatial features within $\{y_i^{\lambda,s}\}_{i=1}^G$.

Finally, the outputs of channel-wise groups and spatial-wise groups are concatenated and reshaped back to the dimensions of the coarse mask y^o, culminate in $y_S^\mu \in \mathbb{R}^{C\times W\times H\times D}$, which is redistributed spatially to further enhance specific axial spatio-temporal feature representation.

Spatial redistribution. To further extract and enhance spatial feature representation from the disentangled segmentation map of the W-axis, we employ a spatial redistribution approach to capture common features in different spatial states.

Methods	Params	FLOPs	FLARE2021			OIMHS			SegTHOR		
			mIoU	Dice	HD95	mIoU	Dice	HD95	mIoU	Dice	HD95
3D U-Net (2016)	5.75	135.88	87.92	93.08	16.31	86.60	92.49	3.40	78.69	87.59	4.32
+PaR	**5.86**	**141.85**	**89.20 +1.28**	**93.89 +0.81**	**2.49 -13.82**	**87.55 +0.95**	**93.08 +0.59**	**2.89 -0.51**	**82.08 +3.39**	**89.82 +2.23**	**2.98 -1.34**
V-Net (2016)	45.61	333.10	84.13	89.89	12.93	81.00	88.53	18.13	75.17	85.12	16.10
+PaR	**45.72**	**337.26**	**85.99 +1.86**	**91.49 +1.60**	**5.98 -6.95**	**83.21 +2.21**	**90.26 +1.73**	**13.52 -4.61**	**76.23 +1.06**	**85.81 +0.69**	**11.49 -4.61**
RAUNet (2019)	70.69	366.89	87.93	93.08	27.37	84.52	91.14	13.61	79.34	88.13	14.58
+PaR	**70.80**	**373.06**	**88.32 +0.39**	**93.28 +0.20**	**26.71 -0.66**	**86.20 +1.68**	**92.25 +1.11**	**5.37 -8.24**	**81.04 +1.70**	**89.08 +0.95**	**2.99 -11.59**
ResUNet (2019)	27.22	902.04	87.38	92.56	30.15	84.06	90.84	3.92	79.61	88.26	3.22
+PaR	**27.33**	**908.01**	**87.83 +0.45**	**92.94 +0.38**	**11.80 -18.35**	**87.07 +3.01**	**92.79 +1.95**	**3.33 -0.59**	**80.23 +0.62**	**88.71 +0.45**	**3.14 -0.08**
SegResNet (2019)	4.70	61.32	86.24	91.81	3.22	83.59	90.52	12.05	77.78	86.99	3.38
+PaR	**4.81**	**67.29**	**88.02 +1.78**	**93.13 +1.32**	**2.78 -0.44**	**84.82 +1.23**	**91.35 +0.83**	**5.07 -6.98**	**81.77 +3.99**	**89.60 +2.61**	**2.87 -0.51**
MultiResUNet (2020)	18.65	324.14	85.92	91.35	9.04	86.53	92.44	3.23	79.87	88.53	26.75
+PaR	**18.76**	**330.11**	**86.76 +0.84**	**91.85 +0.50**	**3.68 -5.36**	**88.22 +1.69**	**93.49 +1.05**	**2.85 -0.38**	**80.81 +0.94**	**89.10 +0.57**	**11.06 -15.69**
UNETR (2021)	92.62	82.58	84.74	90.70	4.63	81.52	89.05	29.15	73.76	84.03	4.71
+PaR	**92.73**	**88.55**	**85.87 +1.13**	**91.48 +0.78**	**3.64 -0.99**	**83.73 +2.21**	**90.55 +1.50**	**7.24 -21.91**	**74.13 +0.37**	**84.18 +0.15**	**4.65 -0.06**
Swin UNETR (2021)	61.99	329.46	88.28	93.23	3.25	87.11	92.82	5.21	78.19	87.26	3.87
+PaR	**62.10**	**335.63**	**89.47 +1.19**	**94.04 +0.81**	**2.61 -0.64**	**87.92 +0.81**	**93.27 +0.45**	**2.92 -2.29**	**78.51 +0.32**	**87.42 +0.16**	**3.62 -0.25**
TransBTS (2021)	30.62	110.12	87.63	92.84	3.54	79.40	87.39	33.52	77.70	86.88	3.84
+PaR	**30.74**	**116.29**	**88.36 +0.73**	**93.27 +0.43**	**2.90 -0.64**	**83.68 +4.28**	**90.55 +3.16**	**21.00 -12.52**	**81.02 +3.32**	**89.13 +2.25**	**3.75 -0.09**
nnFormer (2023)	149.1	224.36	85.50	91.43	5.41	80.54	88.29	25.32	77.27	86.65	5.11
+PaR	**149.21**	**230.53**	**88.83 +3.33**	**93.69 +2.26**	**2.35 -3.06**	**85.50 +4.96**	**91.80 +3.51**	**7.36 -17.96**	**79.00 +1.73**	**87.69 +1.04**	**3.51 -1.60**
3D UX-NET (2023)	53.00	627.90	88.40	93.31	8.85	87.45	93.01	4.61	78.30	87.34	4.69
+PaR	**53.11**	**637.93**	**89.24 +0.84**	**93.84 +0.53**	**2.43 -6.42**	**88.51 +1.06**	**93.66 +0.65**	**2.65 -1.96**	**79.01 +0.71**	**87.77 +0.43**	**3.61 -1.08**

Table 1. The benchmarking results on the FLARE2021, OIMHS, and SegTHOR datasets validate the performance transformation of 11 previous volumetric segmentation methods integrating PaR post-decoder. The metrics for all models with integrated PaR are highlighted, and the performance differences between models with and without PaR integration are highlighted in bold.

The map y_S^{μ} is firstly reshaped into $y_S^{\mu_1} \in \mathbb{R}^{C \times G \times \frac{W}{G} \times H \times D}$, followed by a dimensional exchange between G and $\frac{W}{G}$ dimensions to generate $y_S^{\mu_2} \in \mathbb{R}^{C \times \frac{W}{G} \times G \times H \times D}$. This transformed tensor is then reshaped back to $y_S^{\mu_3} \in \mathbb{R}^{C \times W \times H \times D}$ and added to y_S^{μ} to generate $y_S^{r} \in \mathbb{R}^{C \times W \times H \times D}$. Through the supplement of spatial information and learning unified features across different axes, spatial redistribution augments the model's capacity to comprehend axial spatial dynamics, aiding in the capture of complex spatio-temporal relationships.

4 Experiments

4.1 Datasets and Implementation Details

The proposed PaR module is jointly trained end-to-end with the baseline model, using the 3D feature maps output by the baseline decoder as input to the module. To substantiate the efficacy of our method, we conducted experiments across three publicly available datasets: FLARE2021 [Ma *et al.*, 2022], OIMHS [Ye *et al.*, 2023], and SegTHOR [Lambert *et al.*, 2020]. For all datasets, we employ an 8:1:1 random split for the training, validation, and testing sets. To ensure fairness, we employ identical data preprocessing protocols and hyperparameter configurations. Across all training sessions, we utilize $\mathcal{L}_{DiceCE}$ as the loss function and the AdamW optimizer with a learning rate of 0.0001, over 80,000 iterations, and a batch size of 2. Validation employs a sliding window approach with a 0.5 overlap. Further details on the datasets and experimental setups are provided in the **Appendix**.

4.2 Benchmarking Results

We evaluate the computational cost and performance implications of integrating the PaR into 11 baselines on FLARE2021, OIMHS, and SegTHOR datasets, as demonstrated in quantitative results Table 1 and qualitative results Fig. 3.

FLARE2021. In our evaluation of the FLARE2021 dataset, all models exhibit performance improvements across all metrics, including mIoU, Dice, and HD95. The enhancements in mIoU and Dice range from 0.39% to 3.33% and 0.2% to 2.26%, respectively, underscoring PaR's efficacy in refining global segmentation accuracy through multi-head attention on different spatial configurations along the specific axis. Furthermore, the HD95 metric, which evaluates boundary segmentation accuracy, indicated notable improvements, exemplified by the reduction in HD95 for the 3D U-Net model from 16.31 to 2.49. This demonstrates PaR's ability to correct significant segmentation errors by learning spatial dependencies along the axial dimension.

OIMHS. The benchmark results on the OIMHS dataset further demonstrate robust performance enhancements across all baseline models when integrated with PaR. The mIoU and Dice improve range from 0.81% to 4.96% and 0.45% to 3.51%, respectively. Notably, the UNETR model shows a reduction in HD95 from 29.15 to 7.24, illustrating PaR's effectiveness in correcting boundary delineations through centralized axial disentanglement of volumetric information. Even state-of-the-art models like 3D UX-NET experienced improvements, with mIoU and Dice increasing by 1.06% and 0.65%, respectively. By disentangling the specific axis, PaR can extract and enhance spatio-temporal feature representation to surpass the limitations of the isotropic nature of 3D conv and improve performance.

SegTHOR. The effectiveness of PaR in disentangling time-axial features and enhancing spatio-temporal representation is similarly demonstrated in the SegTHOR dataset, with a stable increase in mIoU and Dice scores by 0.32% to 3.99% and 0.15% to 2.61% across all benchmark models. The boundary correction effects of PaR are also apparent, with reductions in HD95 of the RAUNet model from 14.58 to 2.99. These

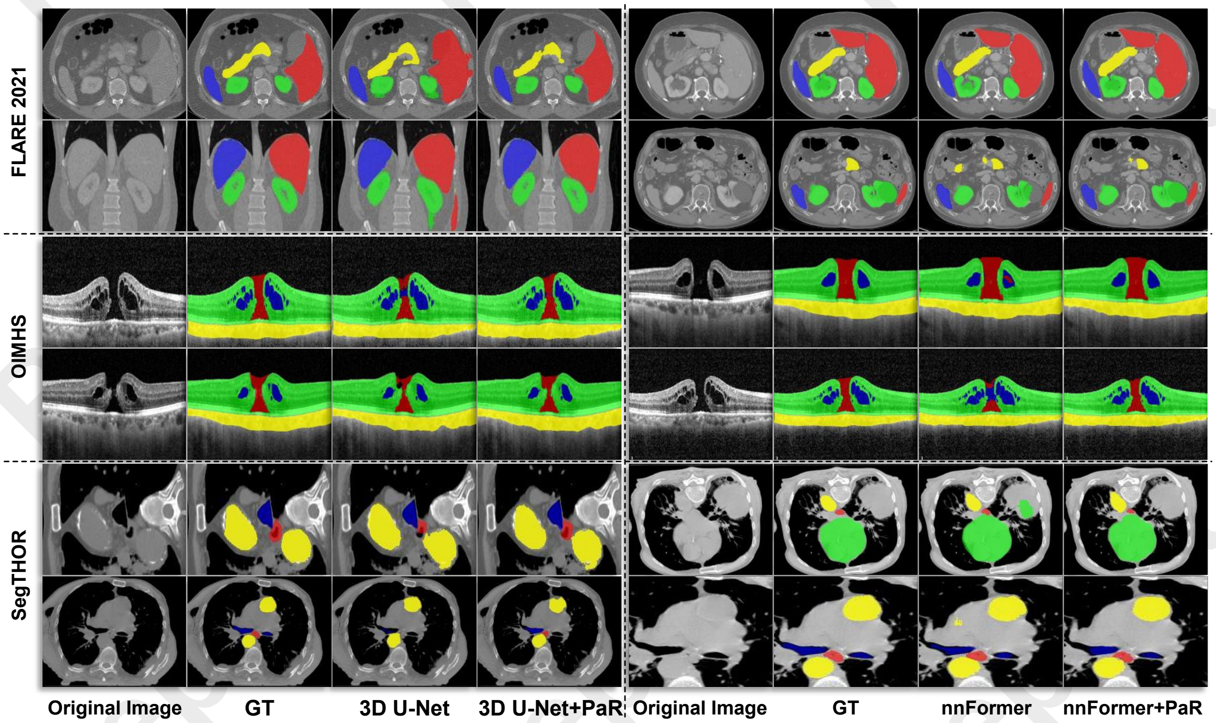

Figure 3: Qualitative validation results of performance transformations using PaR are compared across three public datasets: FLARE2021, OIMHS, and SegTHOR, involving 3D U-Net, 3D U-Net+PaR, nnFormer, and nnFormer+PaR. Segmentation results are shown in different colors for each category. For enhanced visual clarity, the displayed images have been cropped. Please kindly zoom in for a better view.

Method	mIoU	Dice	HD95
3D U-Net	87.92	93.08	16.31
+PDI	88.92	93.73	10.72
+SAI	88.45	93.49	20.33
+PaR-cat	88.67	93.56	8.24
+PaR-mult	88.58	93.48	13.37
+PaR-add	**89.20**	**93.89**	**2.49**

Table 2. Ablation study of key components and fusion strategies within the PaR module on the FLARE2021 dataset.

Method	mIoU	Dice	HD95
UNETR	81.52	89.05	29.15
+PDI	82.88	89.98	22.59
+SAI	81.88	89.21	13.35
+PaR-cat	82.61	89.77	16.87
+PaR-mult	82.07	89.42	27.31
+PaR-add	**83.73**	**90.55**	**7.24**

Table 3. Ablation study of key components and fusion strategies within the PaR module on the OIMHS dataset.

results consistently demonstrate PaR's ability to improve segmentation performance by volumetric axial disentanglement and focused attention along the specific axis.

Experimental results across three datasets confirm that the integration of PaR can consistently enhance performance across CNN-based segmentation models while introducing only a marginal computational cost. This improvement transcends model architectures and dataset characteristics, indicating that volumetric axial disentanglement along a specific axis could effectively boost overall segmentation accuracy and boundary precision. Even in state-of-the-art models, PaR provides measurable gains, validating the robustness and applicability in various 3D medical imaging contexts.

4.3 Ablation Studies

Key components and fusion strategy. To validate the efficacy of our proposed method across CNN-based and Transformer-based architectures, we conducted ablation experiments using 3D U-Net and UNETR as baselines on the FLARE2021 and OIMHS datasets, as shown in Tables 2 and 3. The integration of either the PDI or SAI pipeline individually improved segmentation performance, with PDI capturing and learning specific axial critical spatio-temporal information and SAI applying specific axial feature disentanglement to enhance feature representation. Fusion strategies of PDI and SAI pipelines are also evaluated, including addition, element-wise multiplication, and cat. The addition strategy demonstrates the most obvious improvements across all metrics, underscoring the complementary nature of the PDI and SAI pipelines. Overall, the integrated PaR module consistently outperformed baseline models in segmentation accuracy and boundary precision, validating the effectiveness of specific axial feature disentanglement.

Axial information density reduction. To better understand and validate the efficacy of PaR in W-axial feature disentanglement, we conducted ablation experiments on the FLARE2021 dataset under different frame sampling scenarios, simulating various physical spatial distances in 3D medical imaging, with results shown in Table 4. The experiment

Method	mIoU	Dice	HD95
3D U-Net	87.92	93.08	16.31
+PaR	**89.20 +1.28**	**93.89 +0.81**	**2.49 -13.82**
3D U-Net(1/3)	86.27	91.76	14.52
+PaR	**87.55 +1.28**	**92.64 +0.88**	**8.55 -5.97**
3D U-Net(1/2)	83.25	89.30	16.28
+PaR	**85.29 +2.04**	**90.80 +1.50**	**7.10 -9.18**

Table 4. Ablation study of different frame sampling scenarios.

Method	mIoU	Dice	HD95
3D U-Net	87.92	93.08	16.31
+PaR (W)	**89.20 +1.28**	**93.89 +0.81**	**2.49 -13.82**
+PaR (H)	88.50 +0.58	93.47 +0.39	8.16 -8.15
+PaR (D)	89.10 +1.18	93.78 +0.70	**2.49 -13.82**

Table 5. Ablation study on disentanglement along different axes on the FLARE2021 dataset.

Method	mIoU	Dice	HD95
3D U-Net	86.60	92.49	3.401
+PaR (W)	**87.55 +0.95**	**93.08 +0.59**	**2.89 -0.51**
+PaR (H)	86.96 +0.36	92.72 +0.23	5.12 +1.72
+PaR (D)	87.44 +0.84	93.01 +0.52	2.92 -0.48

Table 6. Ablation study on disentanglement along different axes on the OIMHS dataset.

Heads	mIoU	Dice	HD95
4	88.62	93.44	4.19
8	**89.20**	**93.89**	**2.49**
16	88.56	93.42	2.63
24	88.34	93.34	4.75

Table 7. Ablation study of self-attention heads in PDI.

Ks	mIoU	Dice	HD95
3	**89.20**	**93.89**	5.06
5	88.98	93.73	10.47
7	**89.20**	**93.89**	**2.49**
9	89.05	93.78	5.35
13	89.06	93.77	5.28

Table 8. Ablation study on kernel sizes of SAI's preliminary feature extraction layer.

evaluates the model's performance improvements with PaR integration by deleting every second and third frame to assess the impact of varying time-axial information density. The findings indicate that as time-axial information is reduced, the segmentation performance gains from PaR become more pronounced. This demonstrates that baseline models tend to allocate equal attention to all axes, whereas PaR effectively disentangles information along the W-axis, enhancing the baseline model's ability to extract meaningful spatio-temporal features in scenarios with limited axial data. These results underscore the importance of axial disentanglement in volumetric data when time-axial information density is insufficient.

Disentanglement along different axes. We conduct an ablation study to assess the impact of integrating the PaR module for disentangling along different axes on segmentation performance, validated on the FLARE2021 and OIMHS datasets, with results presented in Table 5 and Table 6. The results demonstrate that disentangling along the W, H, or D axes consistently improves the baseline model's segmentation performance, supporting the hypothesis that specific axial disentanglement mitigates the inherent anisotropic limitations of 3D convs in volumetric data.

Heads impact on axial expression. We conducted ablation experiments on the FLARE2021 dataset to evaluate the impact of varying the number of multi-head self-attention heads in the PDI pipeline, as shown in Table 7, to confirm optimal hyperparameter selection. The results suggest that appropriate heads should be selected based on axial information density, aiming to optimize feature mapping across multiple subspaces, thereby reducing computational overhead and enhancing feature expressiveness. Experiments with 4, 8, 16, and 24 heads reveal that 8 heads provide the best performance among all configurations.

Kernel size effects on feature extraction. An ablation study was conducted on the kernel sizes of SAI's preliminary feature extraction layer on FLARE2021 to determine the optimal hyperparameter settings, with results presented in Table 8. Larger kernel sizes cover a broader local area, capturing more contextual information for edge region recognition. However, excessively large kernels might lead to oversmoothing, losing essential details and texture information. Experiments with kernel sizes of 3 and 7 show close Dice scores, yet the HD95 metric of kernel size 3 is better than that of kernel size 7, indicating superior boundary optimization.

5 Conclusion and Future Work

In this paper, we propose a volumetric axial disentanglement approach to address the pronounced disparities in spatial information across different axial dimensions inherent in volumetric medical imaging. By designing the Post-Axial Refiner (PaR) module, we have enabled a disentangled representation of spatial states and time-axial features, which effectively enhances both the global segmentation performance and the precision of boundaries across several public datasets, including FLARE2021, OIMHS, and SegTHOR. The experimental results underscore the critical importance of axial feature disentanglement in the enhancement of volumetric medical image segmentation, and it only introduces marginal computational overhead. Moreover, the modular design permits adaptation to other volumetric data via axis selection.

The significance of this work lies in the proposal of a concise yet effective construct to address the unique challenge of axial information variability in volumetric medical imaging–a challenge that CNN-based models fail to address with their uniform treatment of all dimensions. Looking ahead, our research will focus on refining the proposed PaR module to enhance its efficiency and lightweight design. This continued research is expected to lead to further enhancements in the accuracy of volumetric segmentation, pushing the performance frontier of 3D conv from a fundamental perspective, and contributing to more precise diagnostic capabilities and treatment planning.

Contribution Statement

Xingru Huang, Jian Huang, and Yihao Guo contributed equally to this work.

References

[Bortsova *et al.*, 2019] Gerda Bortsova, Florian Dubost, Laurens Hogeweg, Ioannis Katramados, and Marleen De Bruijne. Semi-supervised medical image segmentation via learning consistency under transformations. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22*, pages 810–818. Springer, 2019.

[Chen *et al.*, 2021] Jiancheng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

[Chen *et al.*, 2022] Shuai Chen, Zahra Sedghi Gamechi, Florian Dubost, Gijs van Tulder, and Marleen de Bruijne. An end-to-end approach to segmentation in medical images with cnn and posterior-crf. *Medical image analysis*, 76:102311, 2022.

[Dosovitskiy *et al.*, 2020] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[Duan *et al.*, 2019] Jinming Duan, Ghalib Bello, Jo Schlemper, Wenjia Bai, Timothy JW Dawes, Carlo Biffi, Antonio de Marvao, Georgia Doumoud, Declan P O'Regan, and Daniel Rueckert. 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, 2019.

[Goodfellow *et al.*, 2020] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[Hatamizadeh *et al.*, 2021a] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger Roth, Daguang Xu, and Ling Yang. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 574–584, 2021.

[Hatamizadeh *et al.*, 2021b] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI Brainlesion Workshop*, pages 272–284. Springer, 2021.

[He *et al.*, 2017] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[Huang *et al.*,] Xingru Huang, Yihao Guo, Jian Huang, Tianyun Zhang, HE HONG, Shaowei Jiang, and Yaoqi Sun. Upping the game: How 2d u-net skip connections flip 3d segmentation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

[Huang *et al.*, 2022] Wei Huang, Chang Chen, Zhiwei Xiong, Yueyi Zhang, Xuejin Chen, Xiaoyan Sun, and Feng Wu. Semi-supervised neuron segmentation via reinforced consistency learning. *IEEE Transactions on Medical Imaging*, 41(11):3016–3028, 2022.

[Huang *et al.*, 2023] Xingru Huang, Retesh Bajaj, Yilong Li, Xin Ye, Ji Lin, Francesca Pugliese, Anantharaman Ramasamy, Yue Gu, Yaqi Wang, Ryo Torii, et al. Post-ivus: A perceptual organisation-aware selective transformer framework for intravascular ultrasound segmentation. *Medical Image Analysis*, 89:102922, 2023.

[Huang *et al.*, 2024] Xingru Huang, Jian Huang, Kai Zhao, Tianyun Zhang, Zhi Li, Changpeng Yue, Wenhao Chen, Ruihao Wang, Xuanbin Chen, Qianni Zhang, et al. Sasan: Spectrum-axial spatial approach networks for medical image segmentation. *IEEE Transactions on Medical Imaging*, 2024.

[Ibtehaz and Rahman, 2020] Nabil Ibtehaz and M. Sohel Rahman. Multiresunet: Rethinking the u-net architecture for multimodal biomedical image segmentation. *Neural Networks*, 121:74–87, 2020.

[Isensee *et al.*, 2021] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2021.

[Jiang *et al.*, 2022] Yuncheng Jiang, Zixun Zhang, Shixi Qin, Yao Guo, Zhen Li, and Shuguang Cui. Apaunet: axis projection attention unet for small target in 3d medical segmentation. In *Proceedings of the Asian Conference on Computer Vision*, pages 283–298, 2022.

[Kaissis *et al.*, 2021] Georgios Kaissis, Markus R Makowski, Daniel Rückert, and Rickmer F Braren. End-to-end privacy preserving deep learning on multi-institutional medical imaging. *Nature Machine Intelligence*, 3(6):473–484, 2021.

[Ke *et al.*, 2022] Lei Ke, Martin Danelljan, Xia Li, Yu-Wing Tai, Chi-Keung Tang, and Fisher Yu. Mask transfiner for high-quality instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4412–4421, 2022.

[Kirillov *et al.*, 2020] Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross Girshick. Pointrend: Image segmentation as rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9799–9808, 2020.

[Kwon *et al.*, 2019] Gihyun Kwon, Chihye Han, and Daeshik Kim. Generation of 3d brain mri using auto-encoding generative adversarial networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 118–126. Springer, 2019.

[Lambert *et al.*, 2020] Zoé Lambert, Caroline Petitjean, Bernard Dubray, and Su Kuan. Segthor: Segmentation of thoracic organs at risk in ct images. In *2020 Tenth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pages 1–6. IEEE, 2020.

[Lee *et al.*, 2023] Ho Hin Lee, Shunxing Bao, Yuankai Huo, and Bennett A. Landman. 3d UX-net: A large kernel volumetric convnet modernizing hierarchical transformer for medical image segmentation. *The Eleventh International Conference on Learning Representations*, 2023.

[Liao *et al.*, 2020] Xuan Liao, Wenhao Li, Qisen Xu, Xiangfeng Wang, Bo Jin, Xiaoyun Zhang, Yanfeng Wang, and Ya Zhang. Iteratively-refined interactive 3d medical image segmentation with multi-agent reinforcement learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9394–9402, 2020.

[Ma *et al.*, 2022] Jun Ma, Yao Zhang, Song Gu, Xingle An, Zhihe Wang, Cheng Ge, Congcong Wang, Fan Zhang, Yu Wang, Yinan Xu, et al. Fast and low-gpu-memory abdomen ct organ segmentation: the flare challenge. *Medical Image Analysis*, 82:102616, 2022.

[Miao *et al.*, 2023] Jiaxu Miao, Zongxin Yang, Leilei Fan, and Yi Yang. Fedseg: Class-heterogeneous federated learning for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8042–8052, 2023.

[Milletari *et al.*, 2016] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 565–571. IEEE, 2016.

[Myronenko, 2019] Andriy Myronenko. 3d mri brain tumor segmentation using autoencoder regularization. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part II 4*, pages 311–320. Springer, 2019.

[Ni *et al.*, 2019] Zhen-Liang Ni, Gui-Bin Bian, Xiao-Hu Zhou, Zeng-Guang Hou, Xiao-Liang Xie, Chen Wang, Yan-Jie Zhou, Rui-Qi Li, and Zhen Li. Raunet: Residual attention u-net for semantic segmentation of cataract surgical instruments. In Tom Gedeon, Kok Wai Wong, and Minho Lee, editors, *Neural Information Processing*, pages 139–149, Cham, 2019. Springer International Publishing.

[Shaker *et al.*, 2024] Abdelrahman M Shaker, Muhammad Maaz, Hanoona Rasheed, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Unetr++: delving into efficient and accurate 3d medical image segmentation. *IEEE Transactions on Medical Imaging*, 2024.

[Shu *et al.*, 2024] Zhenyu Shu, Teng Wu, Jiajun Shen, Shiqing Xin, and Ligang Liu. Semi-supervised 3d shape segmentation via self refining. *IEEE Transactions on Image Processing*, 2024.

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

[Wang *et al.*, 2021] Wenxuan Wang, Chen Chen, Meng Ding, Hong Yu, Sen Zha, and Jiangyun Li. Transbts: Multimodal brain tumor segmentation using transformer. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24*, pages 109–119. Springer, 2021.

[Xie *et al.*, 2021] Yutong Xie, Jianpeng Zhang, Chunhua Shen, and Yong Xia. Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*, pages 171–180. Springer, 2021.

[Xue *et al.*, 2023] Hengzhi Xue, Qingqing Fang, Yudong Yao, and Yueyang Teng. 3d pet/ct tumor segmentation based on nnu-net with gcn refinement. *Physics in Medicine & Biology*, 68(18):185018, 2023.

[Ye *et al.*, 2023] Xin Ye, Shucheng He, Xiaxing Zhong, Jiafeng Yu, Shangchao Yang, Yingjiao Shen, Yiqi Chen, Yaqi Wang, Xingru Huang, and Lijun Shen. Oimhs: An optical coherence tomography image dataset based on macular hole manual segmentation. *Scientific Data*, 10(1):769, 2023.

[Yuan *et al.*, 2020] Yuhui Yuan, Jingyi Xie, Xilin Chen, and Jingdong Wang. Segfix: Model-agnostic boundary refinement for segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pages 489–506. Springer, 2020.

[Zhang *et al.*, 2021] Gang Zhang, Xin Lu, Jingru Tan, Jianmin Li, Zhaoxiang Zhang, Quanquan Li, and Xiaolin Hu. Refinemask: Towards high-quality instance segmentation with fine-grained features. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6861–6869, 2021.

[Zhou *et al.*, 2023] Hong-Yu Zhou, Jiansen Guo, Yinghao Zhang, Xiaoguang Han, Lequan Yu, Liansheng Wang, and Yizhou Yu. nnformer: Volumetric medical image segmentation via a 3d transformer. *IEEE Transactions on Image Processing*, pages 4036–4045, 2023.

[Çiçek *et al.*, 2016] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: Learning dense volumetric segmentation from sparse annotation. *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016*, pages 424–432, 2016.