Hybrid Local-Window-Attention–Assisted U-Net Model for Multimodal Medical-Image Segmentation

Jiwon Kim[1][0009−0005−9243−1768], Seyong Jin[2][0009−0004−4537−8581], Yeonwoo Noh[3][0009−0004−3235−2948], Hyeonjoon Moon[4*][0000−0003−4528−6009], Minwoo Lee[5][0000−0001−8474−5744], and Wonjong Noh[6*][0000−0001−5668−0453]

[1] Convergence Engineering for Artificial Intelligence, Sejong University, Seoul, Korea
kimjiwonl5124@gmail.com
[2] Artificial Intelligence, Sejong University, Seoul, Korea
[3] College of Medicine, Gachon University, Incheon, Korea
[4] Computer Sciences and Engineering, Sejong University, Seoul, Korea
hmoon@sejong.ac.kr
[5] Neurology, Hallym University Sacred Heart Hospital, Anyang, Korea
[6] School of Information Science, Hallym University, Chuncheon, Korea
wonjong.noh@hallym.ac.kr

Abstract. Multimodal image segmentation has been gaining significance with the advancement of deep learning and increasing diversity of datasets. Although researchers have been actively exploring multimodal U-Net structures, improvements in the segmentation of fine features in medical images remain limited. In this study, we propose a novel U-Net model based on hybrid local-window attention, for multimodal medical-image segmentation. This study aims to effectively analyze overlapping brain-tumor lesions and extract essential information from different magnetic-resonance-imaging modalities for more precise segmentation. The proposed hybrid local-window–attention mechanism comprises local-window self-attention and cross-attention, disentangled representation learning (DRL), and region-aware contrastive learning (RCL) modules. We apply local-window self-attention for achieving efficiency over global attention, and local-window cross-attention between the encoder and decoder to enhance the modality interaction. The hybrid local-window–attention structure extracts modality-specific features, whereas DRL preserves modality and lesion information. RCL utilizes the contrast loss within the lesions to improve segmentation. We perform comprehensive experiments on the BraTS 2023 and BraTS 2024 datasets and confirm that the proposed model provides enhanced medical-image segmentation performance, compared with U-Net based benchmark models without pre-training.

Keywords: Disentangled Representation Learning · Hybrid Attention · Multimodal · Region-aware Contrastive Learning

* Corresponding authors. Emails: hmoon@sejong.ac.kr, wonjong.noh@hallym.ac.kr

1 Introduction

Medical-image analysis has gained importance with the advent of artificial intelligence, and tasks such as classification, detection, and segmentation have been actively investigated [1,2]. Image segmentation supports clinical diagnosis and treatment planning by distinguishing lesions, organs, and tissue in medical images. This process allows detailed examination and is useful in early disease detection, treatment evaluation, and surgical planning [3,4].

For this purpose, two-dimensional (2D) U-Net–based segmentation [5] has been actively studied in the recent years. The symmetric encoder–decoder structure of U-Net preserves spatial information while extracting high-level features, leading to its widespread use in medical-image segmentation. Various extended models have been proposed to improve its performance. Alom et al. [6] introduced a recurrent residual block to enhance feature reuse and maintain deep-feature representation, thereby improving the segmentation performance. However, the repeated structure of this block increased the computational cost, thereby affecting the training stability. Subsequently, some studies [7,8] proposed attention-based 2D U-Net–based segmentation methods. Oktay et al. [7] introduced self-attention to improve tumor-boundary delineation and noise suppression; however, its computational complexity increased the risk of overfitting and model intricacy. Alom et al. [8] integrated recurrent residual structures, attention mechanisms, and multimodal learning to capture both local and global features. However, this combination increased hardware demands and complicated hyperparameter tuning owing to the interactions between components. Recently, three-dimensional (3D) U-Net has emerged as a standard architectural framework, which extends the 2D U-Net design into 3D medical-imaging tasks.

For example, Isensee et al. [9] proposed nnU-Net, which configured model architecture along with pre and postprocessing steps based on dataset characteristics. This approach enabled performance optimization without additional fine-tuning. While applicable to various imaging domains, including brain magnetic resonance imaging (MRI) and computed tomography (CT), its reliance on an automated pipeline reduced its flexibility for custom modifications and increased the execution time when processing large datasets.

On the other hand, with the increasing diversity of datasets and advancements in deep-learning techniques, the need for multimodal image segmentation capable of processing multiple medical-imaging modalities has increased. To address this challenge, multimodal U-Net structures have been researched [10,11]. However, the existing approaches demonstrate limited improvements in the performance of medical-image segmentation tasks that require precise classification of small lesions and detailed structural features [12].

Existing disentangled representation learning (DRL) approaches focus on representation separation within a single modality, lacking mechanisms to disentangle shared and modality-specific features in a multimodal setting or to integrate with contrastive learning [18]. Similarly, prior region-aware contrastive learning (RCL) methods emphasize global inter-class separation, limiting their

ability to capture region-level distinctions and leverage complementary features across modalities [19].

In this study, we propose a new U-Net model, which incorporates hybrid local-window attention to improve multimodal medical-image segmentation. It consists of local-window self-attention, local-window cross-attention, DRL, and RCL as its main function modules. Here, local-window self-attention extracts spatial and structural features from each modality in the encoder to delineate the shape and boundary of the segmented region. Local-window cross-attention enhances feature exchange between the encoder and decoder, enabling efficient fusion of multimodal data. Also, DRL preserves modality-specific features while separating essential representations for learning and RCL reduces redundant features and refines shared feature representations.

2 Methods

We propose a new U-Net model that performs multimodal segmentation. It comprises a multimodal encoder, local-window self-attention module, DRL module, local-window cross-attention module, RCL module, and decoder. It is presented in Fig. 1.

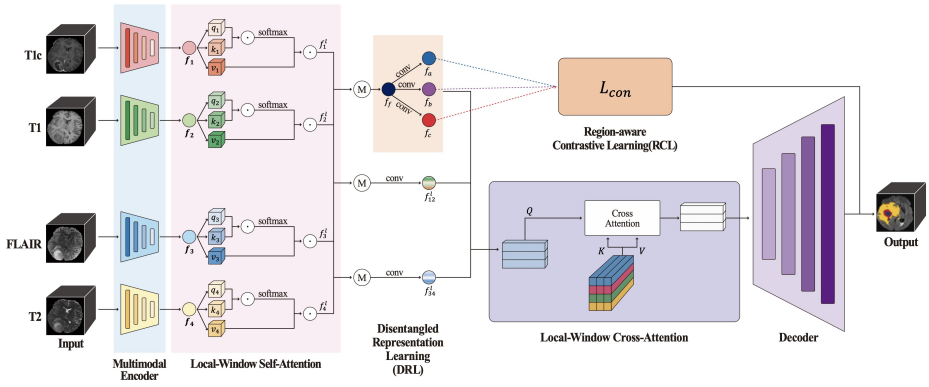

Fig. 1. Overall proposed architecture

2.1 Multimodal Encoder

Each imaging modality (T1, T1w, T2, and FLAIR) was processed using a separate U-Net–based encoder. In the first stage, a DoubleConv3D block with a $3\times3\times3$ convolution, instance normalization (IN), and LeakyReLU expanded the feature channels while maintaining the spatial resolution. In the second to fourth stages, a Conv3D block with a $3\times3\times3$ convolution (stride 2), IN, and LeakyReLU downsampled the features, followed by another DoubleConv3D block. This structure enabled modality-specific feature extraction and hierarchical encoding.

2.2 Local-Window Self-Attention

The input feature map $\mathbf{f} \in \mathbb{R}^{B\times C\times D\times H\times W}$ was divided into 3D windows (patches) of size (w_d, w_h, w_w), where each window contained $w_d \times w_h \times w_w$ tokens (pixels/voxels). A $1\times1\times1$ convolution was independently applied to each partitioned window tensor to obtain the query (Q), key (K), and value (V). The obtained Q, K, V tensors were then rearranged into shapes (batch $\times$ heads $\times$ tokens $\times$ channels), and the attention mechanism was computed as follows:

$$\text{Self-Attention}(Q, K, V) = \text{softmax}\left(\frac{Q \cdot K^\top}{\sqrt{\text{head\_dim}}}\right) \cdot V, \tag{1}$$

where the self-attention results were refined using $1\times1\times1$ convolution to adjust the channels, and the partitioned windows were restored to the original ($D \times H \times W$) spatial structure. Finally, the self-attention results were combined with the original features $\mathbf{f}$ via residual connection to produce the final output.

2.3 Disentangled Representation Learning (DRL)

The multiencoder and local-window self-attention generated four modality-specific feature maps, which were fused into a single feature map f_f using an element-wise mean operation. The DRL block processed f_f through three separate branches (a, b, c), each applying Conv3D filters to extract spatial and channel-wise features. IN and LeakyReLU activation were applied as follows:

$$f_i = \text{LeakyReLU}(\text{IN}(\text{Conv3D}_i(f_f))), \quad i \in \{a, b, c\}. \tag{2}$$

Conv3D filters W^a, W^b, W^c extracted modality-invariant features. IN stabilizes the distributions, and LeakyReLU introduced nonlinearity. The DRL block output three disentangled feature maps:

$$\text{DRL}(f_f) = [f_a, f_b, f_c]. \tag{3}$$

These feature maps enhanced the feature representation in RCL and other tasks.

2.4 Region-aware Contrastive Learning (RCL)

The DRL block generates three 3D feature maps $f_a, f_b, f_c \in \mathbb{R}^{B\times C\times D\times H\times W}$, which contained spatial information and were not directly comparable. To address this, 3D global average pooling (GAP3D) was applied, reducing the spatial dimensions and summarizing the features into vector representations. GAP3D reduced the computational complexity while facilitating feature comparison, converting features into a form suitable for contrastive learning. This transformation is defined as follows:

$$\tilde{f}_a, \tilde{f}_b, \tilde{f}_c \in \mathbb{R}^{B\times C}, \quad \tilde{f}_i = \text{GAP3D}(f_i), \quad i \in \{a, b, c\}. \tag{4}$$

Feature similarity is quantified using the cosine similarity, which measures the alignment between vectors. A higher value indicates a stronger similarity, whereas a lower value indicates a greater dissimilarity. The similarity was computed as follows:

$$\mathrm{sim}(\tilde{f}_a, \tilde{f}_b) = \frac{\tilde{f}_a \cdot \tilde{f}_b}{\|\tilde{f}_a\|\|\tilde{f}_b\|}, \quad \mathrm{sim}(\tilde{f}_a, \tilde{f}_c) = \frac{\tilde{f}_a \cdot \tilde{f}_c}{\|\tilde{f}_a\|\|\tilde{f}_c\|}. \tag{5}$$

The similarity between $\tilde{f}_a$ and $\tilde{f}_b$ was treated as a positive score, indicating proximity to the feature space, whereas the similarity between $\tilde{f}_a$ and $\tilde{f}_c$ was treated as a negative score, indicating dissimilarity. To regulate the similarity scores, a temperature parameter T was applied, which was defined as:

$$\mathrm{pos\_score} = \frac{\mathrm{sim}(\tilde{f}_a, \tilde{f}_b)}{T}, \quad \mathrm{neg\_score} = \frac{\mathrm{sim}(\tilde{f}_a, \tilde{f}_c)}{T}, \tag{6}$$

where a smaller T amplified the contrast between the positive and negative scores, enhancing contrastive learning, whereas a larger T smoothed the similarity distribution. Based on this, the InfoNCE loss was defined as:

$$L_{\mathrm{con}}(a) = -\log\left(\frac{\exp(\mathrm{pos\_score})}{\exp(\mathrm{pos\_score}) + \exp(\mathrm{neg\_score}) + \epsilon}\right), \tag{7}$$

where ϵ denoted a small constant for numerical stability. The loss increased the similarity between the anchor $\tilde{f}_a$ and positive sample $\tilde{f}_b$, while decreasing the similarity between $\tilde{f}_a$ and the negative sample $\tilde{f}_c$. This enabled RCL to improve the feature representation by leveraging contrastive-learning principles.

2.5 Local-Window Cross-Attention & Decoder

The decoder utilized local-window cross-attention to restore the spatial resolution by integrating the encoder feature X with feature maps f_f, f_{12}, and f_{34} obtained from the DRL module. Attention computation was based on key, value, and query definitions. Two sets were considered: encoder-feature–based (K, V, Q) and DRL-feature–based (K', V'). Each feature map was divided into local windows before attention calculation. Scaled dot-product attention was applied as follows:

$$S_w = \frac{Q_w K_w^T}{\sqrt{d_h}}, \quad S'_w = \frac{Q'_w {K'_w}^T}{\sqrt{d_h}}, \quad S''_w = \frac{Q_w f_{12}^T}{\sqrt{d_h}}, \quad S'''_w = \frac{Q_w f_{34}^T}{\sqrt{d_h}}, \tag{8}$$

Softmax normalization converted attention scores into probability distributions, which were used to weigh value vectors and generate the final feature maps. The features computed within the local windows were restored to the original spatial structures by updating the decoder features as follows:

$$G_{l-1} = X_{l-1} + C + C'. \tag{9}$$

The decoder features underwent further refinement through a 1×1 convolution.

$$G_{l-1} = \mathrm{Conv}_{1\times1}(\mathrm{Concat}(X_{l-1}, F_{\mathrm{final}})). \tag{10}$$

Finally, the refined decoder feature G_1 was processed through a 1×1 convolution to generate the segmentation output:

$$Y = W_{\text{out}} G_1. \tag{11}$$

This approach integrated the encoder information and DRL features to enhance the segmentation performance.

2.6 Loss Function

In this study, the overall loss function combined the decoder loss and contrastive loss:

$$L_{\text{seg}} = L_{\text{decoder}} + \alpha L_{\text{con}}, \tag{12}$$

where the weighting factor α controlled the contributions. First, L_{decoder} was a cross-entropy loss that aimed to train the decoder during the final learning stage.

$$L_{\text{decoder}} = \text{CrossEntropy}(Y, Y_{\text{true}}), \tag{13}$$

where Y represented the predicted segmentation map and Y_{true} denoted the ground truth. Next, L_{con} denoted the InfoNCE loss in eq. (7), which trained the RCL [20].

3 Experiments

3.1 Dataset and Implementation Details

Dataset and Preprocessing In this study, we used the publicly available 1,251 BraTS 2023 [13] and 1,350 BraTS 2024 [14] datasets [15]. Each dataset consisted of 3D brain MRI volumes, each of which was divided into four modalities: native (T1), T1-weighted (T1Gd), T2-weighted (T2), and T2 fluid-attenuated inversion recovery (T2-FLAIR). Each MRI volume was $240 \times 240 \times 155$. The data sets were divided into a 7:1:2 ratio for training, validation, and testing, respectively. During training, the four-channel MRI volume centered on the tumor was cropped into a $128 \times 128 \times 128$ voxel patch using linear interpolation and nearest-neighbor interpolation [21]. Z-score normalization was also used to uniformly normalize data with different contrast values. The segmentation targets were classified into three types: Whole Tumor(peritumoral edema, necrosis, non-enhancing tumor, and enhancing tumor, WT), Tumor Core(necrosis and non-enhancing tumor, TC), and Enhancing Tumor(enhancing tumor, ET).

Main Experimental Details and Compared Methods All experiments were performed using Pytorch 2.6.0 and an Nvidia A6000 GPU with 48 GB of memory. The number of epochs was 100, and the batch size was set to 1 [22]. The learning rate was set to 0.0001 (1e-4) [9], and the temperature parameter T was set to 0.07 to enhance RCL [23]. For a baseline comparison, 3D U-Net [5], R2U-Net [6], Attention U-Net [7], R2AU-Net [8], and nnU-Net [9] were used without pre-training.

Evaluation Metrics The performance was evaluated using two metrics: the Dice Score [16] and the Hausdorff Distance 95 (HD95) [17]. These metrics assessed the overlap between segmentation regions and accuracy of the boundary predictions.

3.2 Results and Discussion

This study evaluated the segmentation performance for three tumor regions: WT, TC, and ET. A higher Dice Score indicated better segmentation accuracy, whereas a lower HD95 indicated greater alignment between the predicted tumor boundary and ground truth.

First, we compared the proposed model with U-Net [5], R2U-Net [6], Attention U-Net [7], R2AU-Net [8], and nnU-Net [9] in terms of the Dice Score and HD95 values, the results of which are presented in Table 1. Among the compared models, U-Net [5] exhibited the lowest performance, whereas nnU-Net [9] achieved the highest performance. The proposed model demonstrated the following improvements compared with U-Net [5] and nnU-Net [9]. Compared to U-Net [5], on an average, the proposed model demonstrated a Dice Score increase of 4.19% for WT, 6.51% for TC, 3.97% for ET, and 5.06% in the mean value, while HD95 decreased by 33.60% for WT, 38.47% for TC, 36.04% for ET, and 36.33% in the mean value. Also, compared to nnU-Net [9], on an average, the proposed model demonstrated a Dice Score increase of 1.39% for WT, 3.66% for TC, 1.77% for ET, and 1.96% in the mean value, while HD95 decreased by 24.05% for WT, 26.87% for TC, 34.32% for ET, and 31.14% in the mean value.

Table 1. Comparison of Dice Score and HD95 for U-Net, R2U-Net, Attention U-Net, R2AU-Net, and nnU-Net on BraTS 2023 and BraTS 2024

BraTS 2023: Model Comparison	**Dice Score (%, ↑)**				**HD95 (mm, ↓)**			
	WT	**TC**	**ET**	**Mean**	**WT**	**TC**	**ET**	**Mean**
U-Net [5]	87.82	81.11	78.75	82.56	4.61	6.11	5.97	5.56
R2U-Net [6]	89.24	83.61	79.36	84.07	5.26	5.25	6.00	5.50
Attention U-Net [7]	89.10	83.99	79.57	84.22	4.51	5.03	6.26	5.27
R2AU-Net [8]	89.23	84.46	79.47	84.39	4.17	5.34	6.24	5.25
nnU-Net [9]	90.72	84.08	82.26	85.68	3.78	5.17	6.27	5.07
Proposed	**92.45**	**87.54**	**83.09**	**87.69**	**2.78**	**3.84**	**3.88**	**3.50**
BraTS 2024: Model Comparison	**Dice Score (%, ↑)**				**HD95 (mm, ↓)**			
	WT	**TC**	**ET**	**Mean**	**WT**	**TC**	**ET**	**Mean**
U-Net [5]	88.98	83.27	79.39	83.88	4.00	5.68	6.15	5.28
R2U-Net [6]	89.96	81.92	80.12	84.07	3.60	6.06	6.74	5.47
Attention U-Net [7]	90.26	81.80	80.56	84.20	3.21	5.38	6.01	4.87
R2AU-Net [8]	90.92	83.20	81.64	85.25	3.32	5.09	5.35	4.59
nnU-Net [9]	90.96	84.10	82.50	85.88	3.07	4.70	5.35	4.38
Proposed	**91.74**	**87.51**	**83.11**	**87.45**	**2.90**	**3.42**	**3.87**	**3.40**

Second, we conducted an ablation study on the proposed methods, as listed in Table 2. (Model-1) served as the baseline model, achieving a mean Dice score of 83.74% and mean HD95 of 4.43 mm across both datasets. (Model-2) incorporated self-attention to enhance the spatial-relationship learning, leading to an increase in the Dice scores of TC and ET by 4.89% and 7.79%, respectively, while reducing the mean HD95 to 3.94 mm. (Model-3) applied DRL to improve tumor-boundary segmentation. While it enhanced TC and ET segmentation, the reduction in HD95 remained limited, indicating the need for further refinement. (Model-4) integrated RCL to improve small-lesion recognition. This resulted in a Dice score increase of 7.32% for the ET region and a reduction in the mean HD95 to 3.82 mm, improving boundary accuracy. (Model-5) is fully featured, incorporating cross-attention to enhance feature interaction and information sharing. As a result, the mean Dice Score improved by 4.58% over the baseline, while the mean HD95 decreased by 22.80%, demonstrating the best segmentation performance among all models.

Table 2. Segmentation performance of different methods based on Dice Score and HD95 for BraTS 2023 and BraTS 2024

BraTS2023: Segmentation Methods	**Dice Score (%, ↑)**				**HD95 (mm, ↓)**			
	WT	**TC**	**ET**	**Mean**	**WT**	**TC**	**ET**	**Mean**
(Model-1) Baseline	90.57	83.92	77.26	83.92	3.58	4.57	5.89	4.68
(Model-2) Baseline + Self-Att	90.95	85.05	78.21	84.74	3.17	3.89	4.59	3.88
(Model-3) Baseline + Self-Att + DRL	90.84	85.82	79.33	85.33	3.25	3.86	5.11	4.07
(Model-4) Baseline + Self-Att + DRL + RCL	91.52	86.81	82.37	86.90	3.03	4.29	4.57	3.96
(Model-5) Baseline + Self-Att + DRL + RCL + Cross-Att	**92.45**	**87.54**	**83.09**	**87.69**	**2.78**	**3.84**	**3.88**	**3.50**
BraTS2024: Segmentation Methods	**Dice Score (%, ↑)**				**HD95 (mm, ↓)**			
	WT	**TC**	**ET**	**Mean**	**WT**	**TC**	**ET**	**Mean**
(Model-1) Baseline	90.75	82.97	76.94	83.55	3.54	3.78	5.20	4.17
(Model-2) Baseline + Self-Att	90.26	81.75	80.78	84.26	3.17	4.56	4.24	3.99
(Model-3) Baseline + Self-Att + DRL	90.64	85.87	77.87	84.79	3.35	3.66	5.27	4.10
(Model-4) Baseline + Self-Att + DRL + RCL	91.61	86.62	80.68	86.31	2.80	3.96	4.26	3.68
(Model-5) Baseline + Self-Att + DRL + RCL + Cross-Att	**91.74**	**87.51**	**83.11**	**87.45**	**2.40**	**3.64**	**3.93**	**3.32**

4 Conclusions

This study proposed U-Net–based new framework that integrated hybrid local-window attention, DRL, and RCL for multimodal medical image segmentation. The proposed model was evaluated using the BraTS 2023 and 2024 datasets, with the Dice Score and HD95 as evaluation metrics, and through comparisons with U-Net based benchmark models without pre-training. The proposed model confirmed an improvement in medical-image segmentation performance. Ablation studies verified that the hybrid local-window attention, DRL, and RCL contributed to the model's performance enhancement. The model effectively captured modality-specific overlapping tumor structures, leading to improved

segmentation of the ET, TC, and WT regions. Compared with U-Net variants, the proposed approach extracted more detailed features, resulting in improved segmentation accuracy. As future works, we will improve the efficiency of the proposed attention mechanisms and perform more extensive comparisons with pre-trained models. We will also investigate knowledge distillation and quantization techniques to reduce model size, increase training and inference speed, and improve memory usage.

Acknowledgments. This research was supported by the Bio&Medical Technology Development Program of the National Research Foundation (NRF) funded by the Korean government (MSIT) (No. RS-2023-00223501).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A. W. M., van Ginneken, B., Sánchez, C. I.: A survey on deep learning in medical image analysis. Computers in Biology and Medicine **177**, 60–88 (2017). https://doi.org/10.1016/j.media.2017.07.005
2. Shen, D., Wu, G., Suk, H. I.: Deep learning in medical image analysis. Annual review of biomedical engineering **19**(1), 221–248 (2017). https://doi.org/10.1146/annurev-bioeng-071516-044442
3. Choi, S. Y., Kim, J. H., Chung, H. S., Lim, S., Kim, E. H., Choi, A.: Impact of a deep learning-based brain CT interpretation algorithm on clinical decision-making for intracranial hemorrhage in the emergency department. Scientific Reports **14**(1), 1–10 (2024). https://doi.org/10.1038/s41598-024-73589-0
4. Kundisch, A., Hönning, A., Mutze, S., Kreissl, L., Spohn, F., Lemcke, J., Sitz, M., Sparenberg, P., Goelz, L.: Deep learning algorithm in detecting intracranial hemorrhages on emergency computed tomographies. PLoS One **16**(11), 1–18 (2021). https://doi.org/10.1371/journal.pone.0260560
5. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, LNIP, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
6. Alom, M. Z., Yakopcic, C., Hasan, M., Taha, T. M., Asari, V. K.: Recurrent residual U-Net for medical image segmentation. Journal of medical imaging **6**(1), 1–16 (2019). https://doi.org/10.1117/1.JMI.6.1.014006
7. Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla. Y. N., Kainz, B., Glocker, B., Rueckert, D.: Attention U-Net: Learning Where to Look for the Pancreas. arXiv preprint arXiv:1804.03999 (2018). https://doi.org/10.48550/arXiv.1804.03999
8. Zuo, Q., Chen, S., Wang, Z.: R2AU-Net: attention recurrent residual convolutional neural network for multimodal medical image segmentation. Security and Communication Networks **2021**(1), 1–10 (2021). https://doi.org/10.1155/2021/6625688
9. Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J., Maier-Hein, K. H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021). https://doi.org/10.1038/s41592-020-01008-z

10. Fu, Y., Lei, Y., Wang, T., Curran, W. J., Liu, T., Yang, X.: A review of deep learning based methods for medical image multi-organ segmentation. Physica Medica **85**, 107–122 (2021). https://doi.org/10.1016/j.ejmp.2021.05.003
11. Li, Y., Daho, M. E. H., Conze, P. H., Zeghlache, R., Le Boité, H., Tadayoni, R., Cochener B., Lamard, M., Quellec, G.: A review of deep learning-based information fusion techniques for multimodal medical image classification. Computers in Biology and Medicine **177**, 1–29 (2024). https://doi.org/10.1016/j.compbiomed.2024.108635
12. Fu, Y., Lei, Y., Wang, T., Curran, W. J., Liu, T., Yang, X.: Deep learning in medical image registration: a review. Physics in Medicine & Biology **65**(20), 1–48 (2020). https://doi.org/10.1088/1361-6560/ab843e
13. BraTS 2023 Kaggle Homepage, https://www.kaggle.com/datasets/shakilrana/brats-2023-adult-glioma, last accessed 2023
14. BraTS 2024 Synapse Homepage, https://www.synapse.org/Synapse:syn53708249/wiki/627759, last accessed 2024
15. Bakas, S., Baid, U., Rudie, J., Calabrese, E., Aboian, M., Anazodo, U., Conte, G. M., Albrecht, J., Li, H. B., Kofler, F., et al.: BraTS 2024 Cluster of Challenges (BraTS + Beyond-BraTS). Zenodo (2024). https://doi.org/10.5281/zenodo.10978907
16. Dice, L. R.: Measures of the Amount of Ecologic Association Between Species. Ecology **26**(3), 297–302 (1945). https://doi.org/10.2307/1932409
17. Carré, A., Deutsch, E., Robert, C.: Automatic Brain Tumor Segmentation with a Bridge-Unet Deeply Supervised Enhanced with Downsampling Pooling Combination, Atrous Spatial Pyramid Pooling, Squeeze-and-Excitation and EvoNorm. In: International MICCAI Brainlesion Workshop-BrainLes 2021, LNCS, vol. 12963, pp. 253—266. Springer, Cham (2021). https://doi.org/10.1007/978-3-031-09002-8_23
18. Wang, X., Chen, H., Tang, S., Wu, Z., Zhu, W.: Disentangled Representation Learning. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(12), 9677–9696 (2024). https://doi.org/10.1109/TPAMI.2024.3420937
19. Hu, H., Cui, J., Wang, L.: Region-aware contrastive learning for semantic segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16291–16301. IEEE, ICCV (2021). https://doi.org/10.1109/ICCV48922.2021.01598
20. Oord, A. V. D., Li, Y., Vinyals, O.: Representation Learning with Contrastive Predictive Coding. arXiv preprint arXiv:1807.03748 (2019). https://doi.org/10.48550/arXiv.1807.03748
21. Ferreira, A., Solack, N., Li, J., Dammann, P., Kleesiek, J., Alves, V., Egger J.: How we won BraTS 2023 Adult Glioma challenge? Just faking it! Enhanced Synthetic Data Augmentation and Model Ensemble for brain tumour segmentation. arXiv preprint arXiv:2402.17317 (2024). https://doi.org/10.48550/arXiv.2402.17317
22. Capellán-Martín, D., Jiang, Z., Parida, A., Liu, X., Lam V., Nisar, H., Tapp, A., et al.: Model ensemble for brain tumor segmentation in magnetic resonance imaging. In: International Challenge on Cross-Modality Domain Adaptation for Medical Image Segmentation–CrossMoDA BraTS 2023, LNCS, vol. 14669, pp. 221–232. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-76163-8_20
23. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A Simple Framework for Contrastive Learning of Visual Representations. In: Proceedings of the 37th International Conference on Machine Learning, p.1597–1607. IMLS, ICML (2020). https://dl.acm.org/doi/10.5555/3524938.3525087