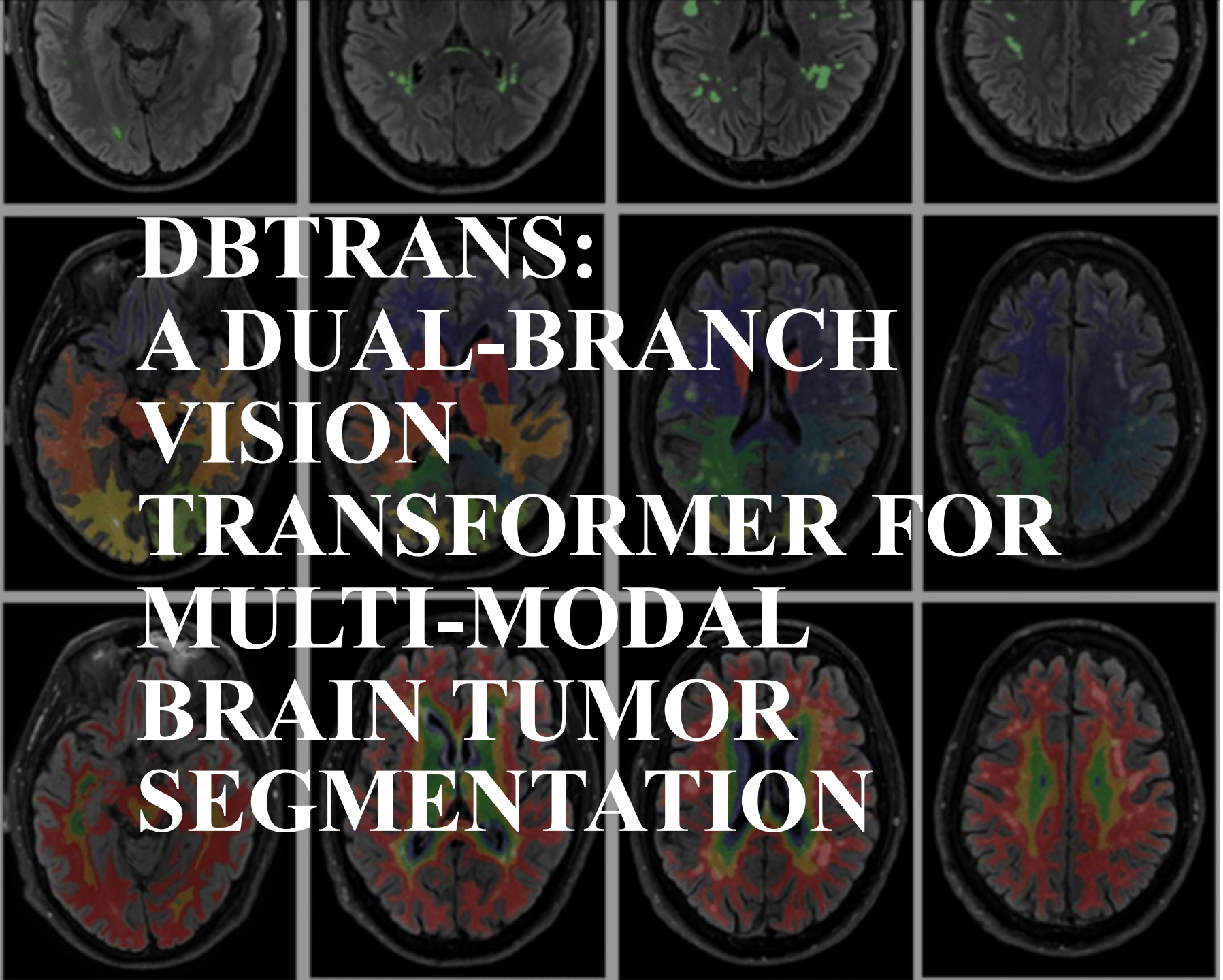


DBTRANS: A DUAL-BRANCH VISION TRANSFORMER FOR MULTI-MODAL BRAIN TUMOR SEGMENTATION

Intelligent Analysis of Biomedical Images

Fall-2023

Seyed Amir Kasaei

Prof. M H Rohban

Introduction

Brain Tumor Semantic segmentation of gliomas based on 3D spatially aligned MRI

- **U-Net based**
  - lack of global understanding of images for convolution operation
  - struggle to model the dependencies between distant features and make full use of the contextual information
- **Transformer**
  - self-attention mechanism (can capture long-range dependencies)
  - When applying 2D models, 3D images need to be sliced along one dimension (may lead to the loss of local information)
  - difficulty capturing global interaction information while effectively encoding local information (just stack modalities (equally treated) and pass them through a network)

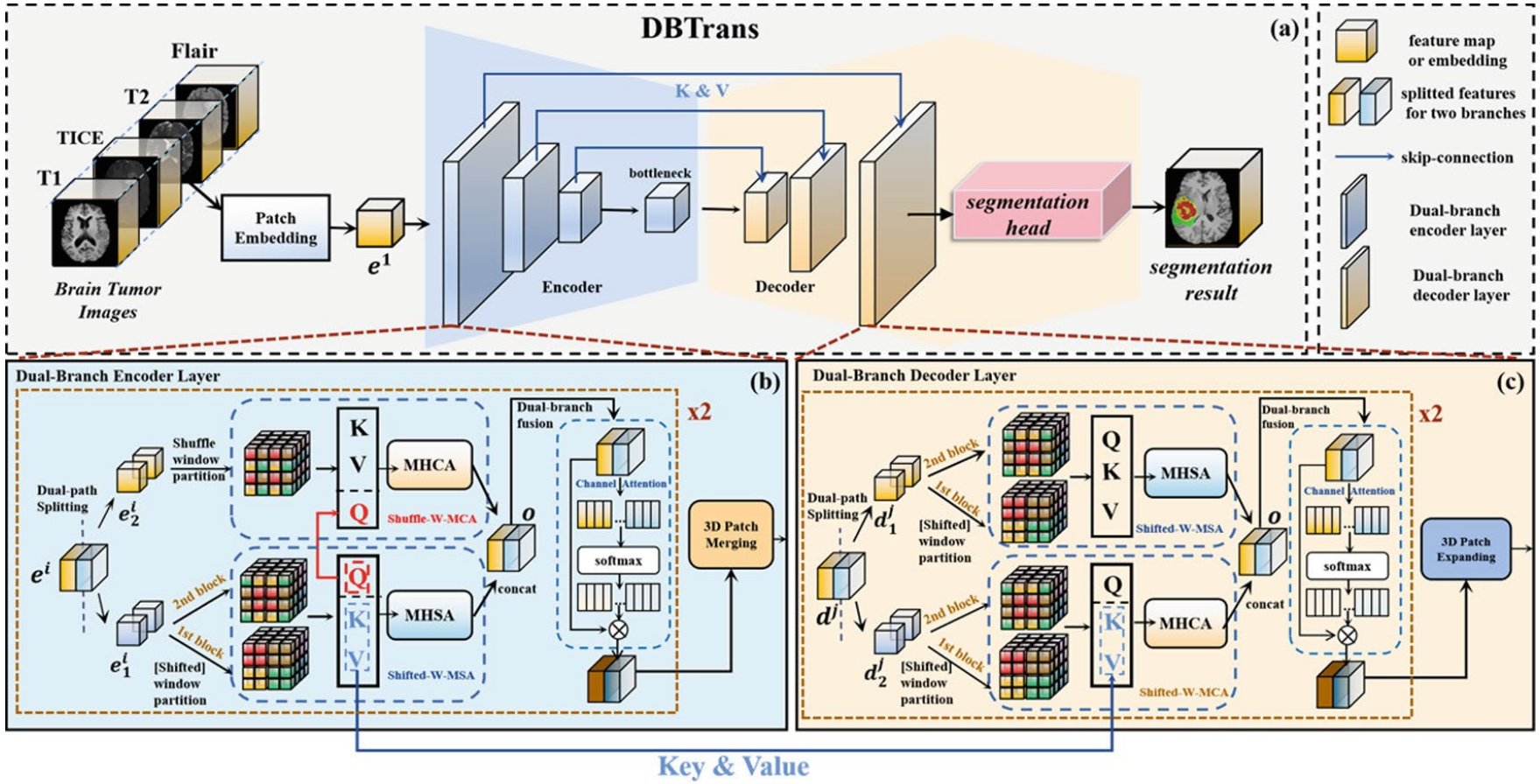
DBTrans

A novel encoder-decoder model for
multi-modal medical image segmentation

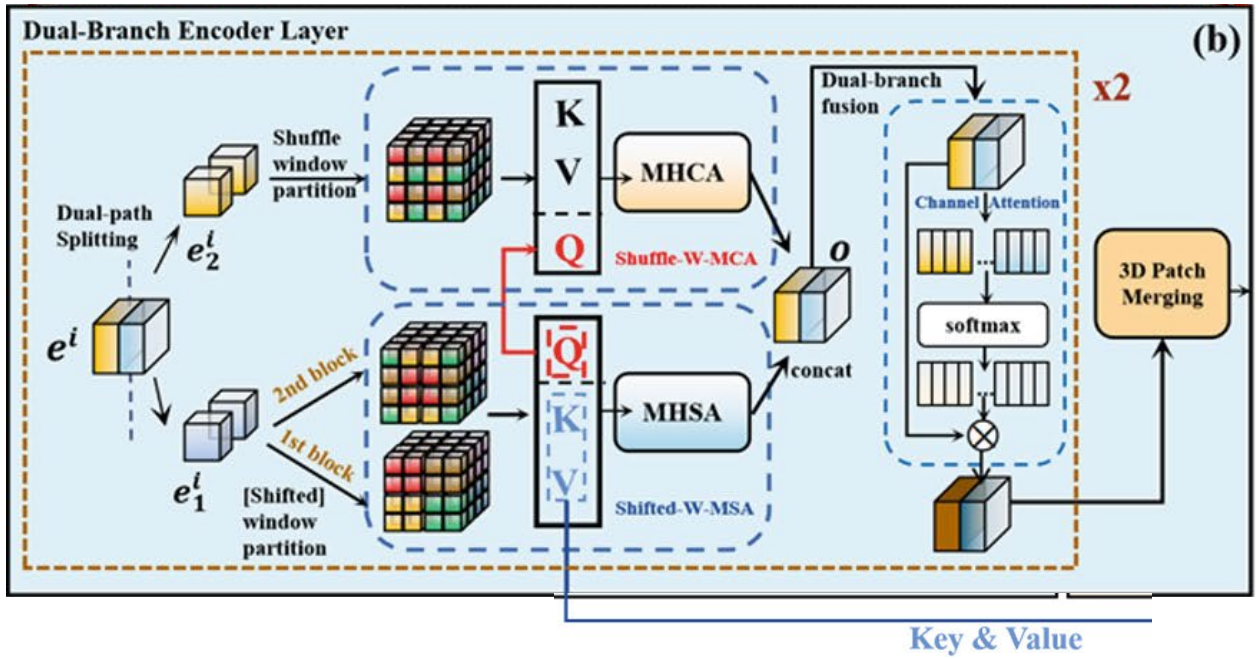
- **Dual-Branch Encoder**
  - two types of window-based attention mechanisms
    - **Shuffled** Window-based Multi-head Cross Attention
    - **Shifted** Window-based Multi-head Self Attention
  - in parallel to dual-branch encoder layers
- **Dual-Branch Decoder**
  - two types of window-based attention mechanisms
    - **Shifted** Window-based Multi-head Cross Attention
    - **Shifted** Window-based Multi-head Self Attention
  - in parallel to dual-branch encoder layers
- **Both local and global feature extraction**
- **Designed for 3D medical images, avoiding the information loss caused by data slicing**

Methodology

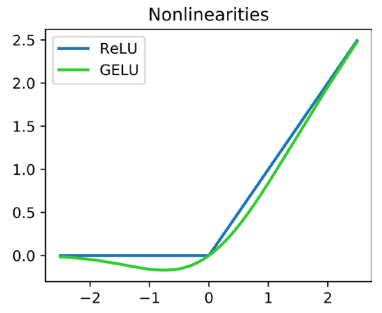
- MRI data of D×H×W ×C with **four** modalities stacked along channel dimensions as the input
- 3D patch embedding ($e^1 \in R^{D1 \times H1 \times W1 \times C1}$)
- U-shaped model (Dual-Branch Encoder-Decoder)
- Segmentation head ($D \times H \times W \times K$)



Architecture
Dual-Branch in Encoder



- Four dual-branch encoder layers including bottleneck
- Each encoder layer consists of two consecutive encoder blocks
- $e^i \in R^{D_i \times H_i \times W_i \times C_i}$ is splitted along the channel dimension $e_1^i, e_2^i \in R^{D_i \times H_i \times W_i \times [C_i/2]}$
- **Shifted W-MSA-Based Local Branch**
  - e_1^i is split into non-overlapping windows after a layer normalization to obtain the window matrix m_1^i (WP)
  - The whole feature map is shifted by half of the window size
  - Apply projection matrices W_Q^i, W_K^i, W_V^i to obtain Q_1^i, K_1^i, V_1^i to get attention score
  - layer normalization (LN)
  - multi-layer perceptron (MLP) with two fully connected layers
  - Gaussian Error Linear Unit (GELU)
  - Residual connection after each module



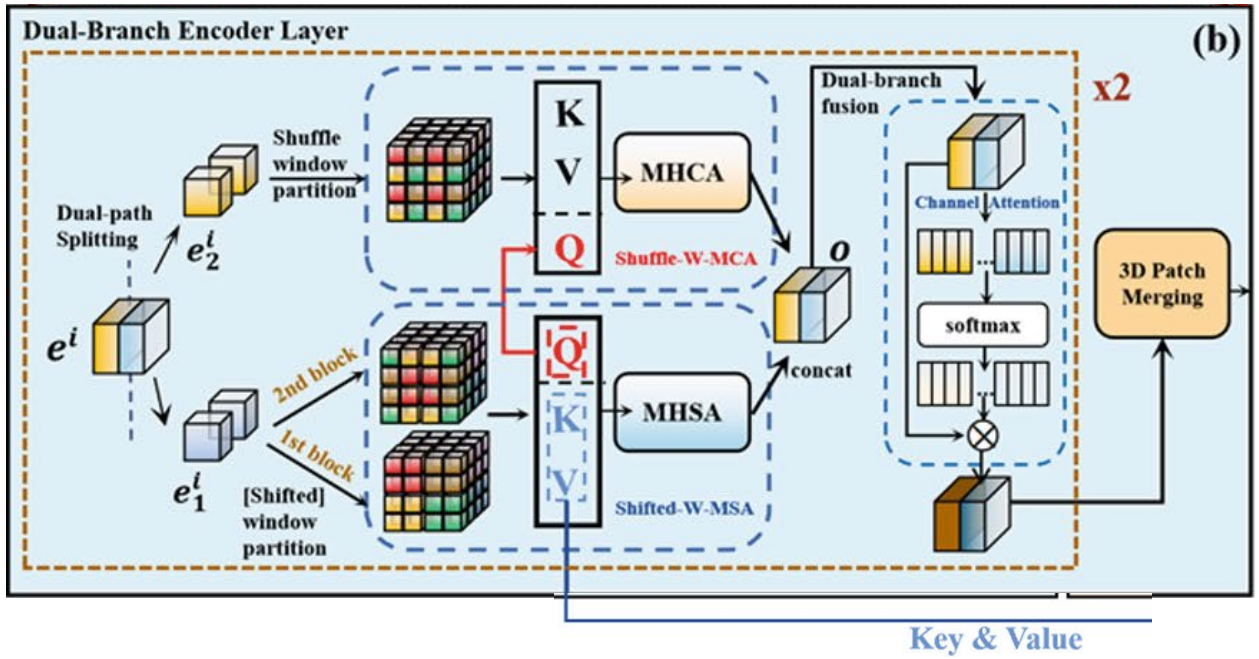
$$m_1^i = [\textbf{\textit{Shifted}}\text{-}]WP\left(LN\left(e_1^i\right)\right),$$

$$Q_1^i, K_1^i, V_1^i \in R^{\frac{D_i H_i W_i}{M^3} \times M^3 \times [C_i/2]} = Proj^i\left(m_1^i\right) = W_Q^i \cdot m_1^i, W_K^i \cdot m_1^i, W_V^i \cdot m_1^i,$$

$$\hat{z}_1^i = W\text{-}MSA\left(Q_1^i, K_1^i, V_1^i\right),$$

$$o_1^i = MLP^i\left(LN\left(\left(\hat{z}_1^i + e_1^i\right)\right)\right) + \left(\hat{z}_1^i + e_1^i\right),$$

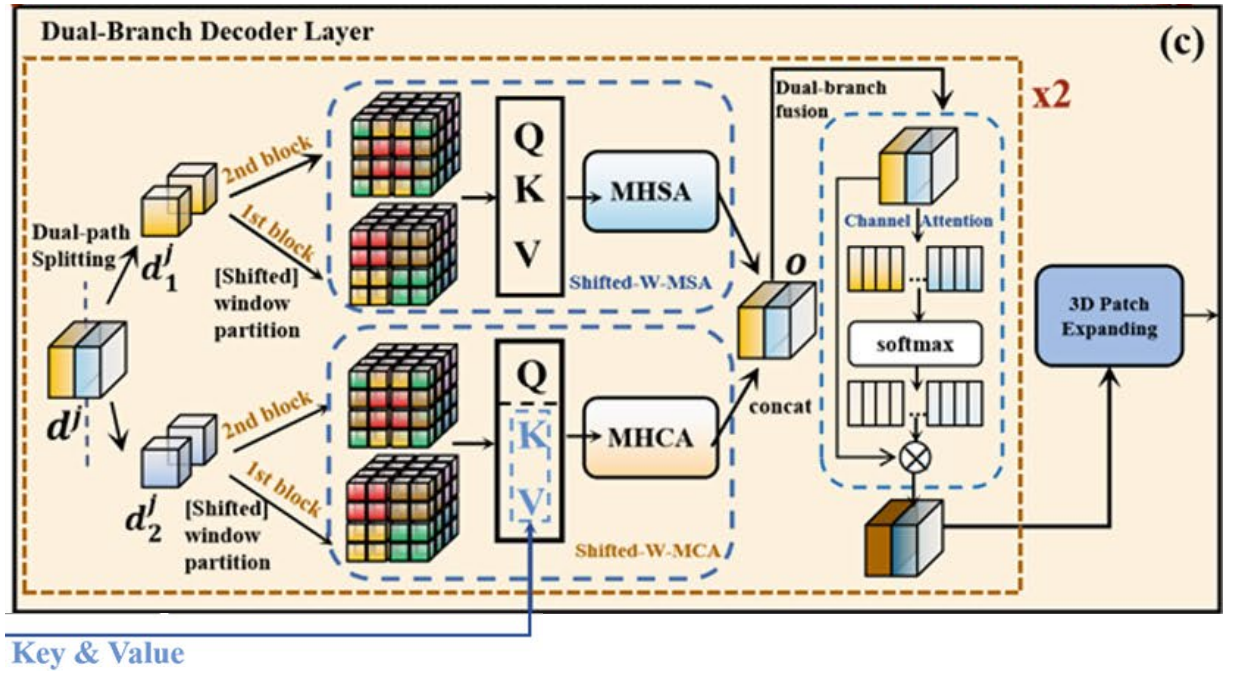
Architecture
Dual-Branch in Encoder



- **Shuffle W-MCA-Based Global Branch**
  - window partition (WP) converts the embedding $e_2^i \in R^{D_i \times H_i \times W_i \times [C_i/2]}$ to $m_2^i \in R^{\frac{(D_i \times H_i \times W_i)}{M^3} \times M^3 \times [C_i/2]}$
  - Shuffle operations on the patches in different windows
    (patches at the same relative position in different windows are rearranged together in a window)
  - The query from the local branch while generating keys and values from m_2^i
  - Compute cross-attention scores

$$
\begin{aligned}
Q_2^i, K_2^i, V_2^i \in R^{\frac{D_i H_i W_i}{M^3} \times M^3 \times [C_i/2]} &= Proj^i\left(Shuffle(m_2^i)\right), \\
\hat{z}_2^i &= W\text{-}MCA\left(Q = Q_1^i, K = K_2^i, V = V_2^i\right), \\
o_2^i &= MLP^i\left(LN\left(\hat{z}_2^i + e_2^i\right)\right) + \left(\hat{z}_1^i + e_1^i\right).
\end{aligned}
$$

Architecture
Dual-Branch in Decoder



- Three dual-branch decoder layers
- Each decoder layer consists of two consecutive decoder blocks
- $d^j \in R^{D_j \times H_j \times W_j \times C_j}$ is divided into the feature maps

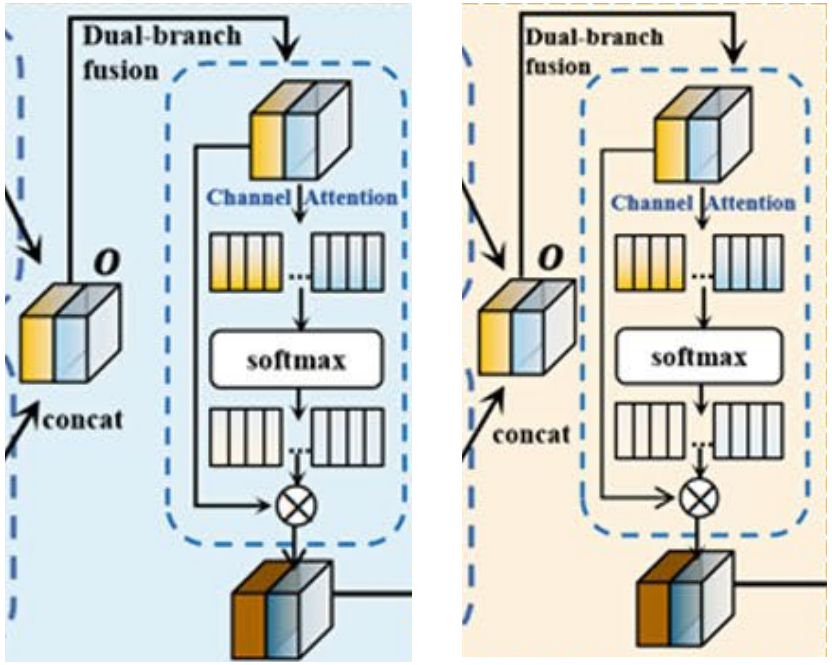
$$d_1^j, d_2^j \in R^{D_j \times H_j \times W_j \times [C_j/2]}$$

- The local branch based on Shifted-W-MSA is the same as that in the encoder
- **Shifted W-MCA-Based Global Branch**
  - global branch receives the **query** matrix from the split feature map
  - Receiving **key** (K_{e^i}) and **value** (V_{e^i}) matrices from the encoder block in the corresponding stage

$$Q_2^j, K_2^j, V_2^j = Proj^j\left(\left[\textbf{\textit{Shifted}}\text{-}\right]WP\left(LN\left(d_2^i\right)\right)\right),$$
$$\hat{z}_2^j = W\text{-}MCA\left(Q = Q_2^j, K = K_{e_1^{4-j}}, V = V_{e_1^{4-j}}\right),$$
$$o_2^j = MLP^i\left(LN\left(\hat{z}_2^j + d_2^j\right)\right) + \left(\hat{z}_1^j + d_1^j\right),$$

Architecture

Channel-Attention-Based Dual-Branch Fusion



- Combines the features $o_1^m, o_2^m \in R^{D_m \times H_m \times W_m \times [C_m/2]}$ for each encoder or decoder layer
- The dependencies between the feature channels within the individual branches are implicitly modeled with the SE-Weight assignment first proposed in **Squeeze-and-excitation networks**
- Dynamically assign weights for both dual-branch fusion and multi-modal fusion
- $Z_p^m \in R^{[C/2] \times 1 \times 1 \times 1}$ is the attention weight of a single branch
- Re-calibrated using a Softmax function
- multiplied with the corresponding scale feature map to
- obtain the refined output feature map with richer multi-scale feature information

$$Z_p = SE\_Weight\left(o_p^m\right), p = 1, 2,$$

$$attn_p = Softmax\left(Z_p\right) = \frac{\exp(Z_p)}{\sum_{p=1}^{2} \exp(Z_p)},$$

$$Y_p = o_p^m \odot attn_p, p = 1, 2,$$

$$O = Cat([Y_1, Y_2]),$$

DBTrans

Contribution

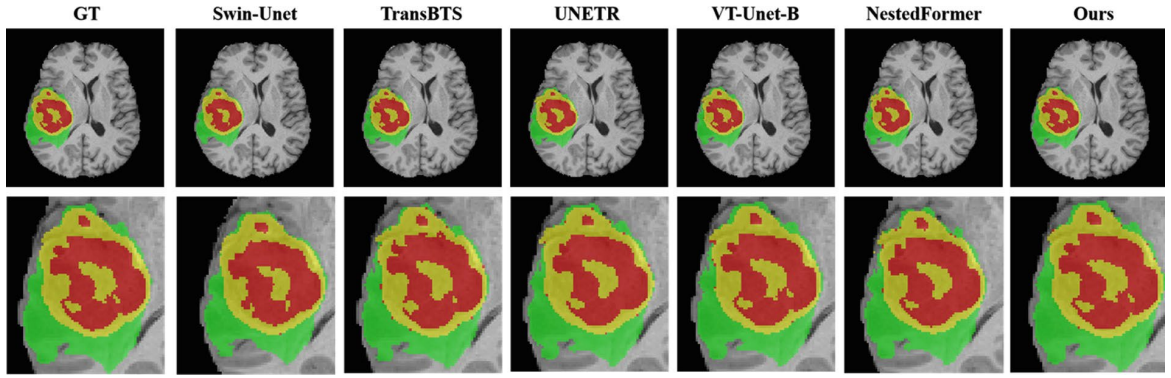
- Two attention mechanisms to model close-window and distant-window dependencies without any extra computational cost.
- An extra path to facilitate the decoding process in addition to traditional skip-connection (Shifted-W-MCA-based global branch – a bridge between encoder and decoder)
- Strengthening the fusion effect of the multi-modal information from a global perspective by improving channel attention

Training Details

- Cross Entropy Loss & Dice Loss
- 300 epochs on a single RTX 3090 with 24G
- Adam optimizer
- Learning rate was set to 1×10−4
- Data augmentation includes random flipping, intensity scaling and intensity shifting on each axis with probabilities set to 0.5, 0.1 and 0.1

Experiments & Results

- **Dataset**
  - Multimodal Brain Tumor Segmentation Challenge (BraTS 2021)
  - Four modalities T1, T1CE, T2, Flair
  - Three distinct sub-regions of brain tumors peritumoral edema, enhancing tumor, and tumor core
- **Comparative Experiments**
  - the Dice scores and 95% Hausdorff distances of different methods for segmentation results
  - Three different tumor regions Enhancing Tumor(ET), Tumor Core(TC) and Whole Tumor(WT)
  - Higher Dice score and a lower 95% HD indicates better performance
  - Floating point operation per second(FLOPS)



Method	#param	FLOPS	Dice Score				95% Hausdorff Distance			
			ET	TC	WT	AVG	ET	TC	WT	AVG
3D U-Net[31]	11.9M	557.9G	83.39	86.28	89.59	86.42	6.15	**6.18**	11.49	7.94
Swin-Unet[27]	52.5M	93.17G	83.34	87.62	89.81	89.61	6.19	6.35	11.53	8.03
TransBTS[25]	33M	333G	80.35	85.35	89.25	84.99	7.83	8.21	15.12	10.41
UNETR[22]	102.5M	193.5G	79.78	83.66	90.10	84.51	9.72	10.01	15.99	11.90
nnFormer[23]	39.7M	110.7G	82.83	86.48	90.37	86.56	8.00	7.89	11.66	9.18
VT-Unet-B[26]	20.8M	165.0G	85.59	87.41	91.02	88.07	6.23	6.29	10.03	7.52
NestedFormer[36]	**10.48M**	**71.77G**	85.62	88.18	90.12	87.88	**6.08**	6.43	10.23	7.63
DBTrans(Ours)	24.6M	146.2G	**86.70**	**90.26**	**92.41**	**89.69**	6.13	6.24	**9.84**	**7.38**

Ablation Study

1) SwinUnet-1 (baseline): We use Swin-Transformer layers without any dual-branch module
2) SwinUnet-2: Based on (1), we add dual-branch encoder layers to the model
3) SwinUnet-3: Based on (1), we add dual-branch decoder layers to the model
4) SwinUnet-4: Based on (1), add both the encoder and decoder without the dual-branch fusion module
5) Our proposed DBTrans

Name	Index	DB-E	DB-D	Dual-branch fusion	Avg-Dice	Param
SwinUnet-1	(1)	×	×	×	86.73	52.5M
SwinUnet-2	(2)	√	×	×	87.52	43.2M
SwinUnet-3	(3)	×	√	×	88.26	46.1M
SwinUnet-4	(4)	√	√	×	88.86	27.7M
proposed	**(5)**	√	√	√	**89.69**	**24.6M**