scientific reports

Check for updates

OPEN

DSIT UNet a dual stream iterative transformer based UNet architecture for segmenting brain tumors from FLAIR MRI images

Shakib Al Hasan[1,2], S. M. Mahim[1,2], Md Emamul Hossen[1,2], Md Olid Hasan[1,2], Md Khairul Islam[1,2]✉, Patrizia Livreri[3]✉, Salah Uddin Khan[4,5], Mohammad Alibakhshikenari[6]✉ & Md Sipon Miah[7,8]✉

Brain tumor segmentation remains challenging in medical imaging with conventional therapies and rehabilitation owing to the complex morphology and heterogeneous nature of tumors. Although convolutional neural networks (CNNs) have advanced medical image segmentation, they struggle with long-range dependencies because of their limited receptive fields. We propose Dual-Stream Iterative Transformer UNet (DSIT-UNet), a novel framework that combines Iterative Transformer (IT) modules with a dual-stream encoder–decoder architecture. Our model incorporates a transformed spatial–hybrid attention optimization (TSHAO) module to enhance multiscale feature interactions and balance local details with the global context. We evaluated DSIT-UNet using three benchmark datasets: The Cancer Imaging Archive (TCIA) from The Cancer Genome Atlas (TCGA), BraTS2020, and BraTS2021. On TCIA, our model achieved a Mean Intersection over Union of 95.21%, mean Dice Coefficient of 96.23%, precision of 95.91%, and recall of 96.55%. BraTS2020 attained a Mean IoU of 95.88%, mDice of 96.32%, precision of 96.21%, and recall of 96.44%, surpassing the performance of the existing methods. The superior results of DSIT-UNet demonstrate its effectiveness in capturing tumor boundaries and improving segmentation robustness through hierarchical attention mechanisms and multiscale feature extraction. This architecture advances automated brain tumor segmentation, with potential applications in clinical neuroimaging and future extensions to 3D volumetric segmentation.

Keywords Brain tumor, Deep learning, Segmentation, CNN, Dual-stream, DSIT-UNet, TSHAO, Rehabilitation, Transformer, Iterative transformer

Brain tumor segmentation remains one of the most challenging tasks in medical image analysis, particularly because of the significant heterogeneity of tumors in terms of their location, shape, and size. Despite the potential impact of automated segmentation methods in clinical settings, the complex nature of brain tumors continues to pose obstacles, particularly in terms of achieving an accurate delineation across various MRI modalities[1]. Gliomas, which can be classified as primary or secondary tumors, vary greatly in terms of their biological behavior and appearance on imaging. While low-grade gliomas (LGGs) grow slowly, high-grade gliomas (HGGs), such as glioblastoma multiforme (GBM), are aggressive, rapidly progressing tumors with a poor prognosis[2]. Additionally, the classification and diagnosis of brain tumors, especially in pediatric populations, remain a challenge because of the diversity in tumor types and locations, with approximately 13% of CNS tumors in children found in the brainstem[3].

[1]Department of Biomedical Engineering, Islamic University, Kushtia 7003, Bangladesh. [2]Bio-Imaging Research Lab, BME, Islamic University, Kushtia 7003, Bangladesh. [3]Department of Engineering, University of Palermo, 90128 Palermo, Italy. [4]Sustainable Energy Technologies Center, College of Engineering, King Saud University, 11421 Riyadh, Saudi Arabia. [5]King Salman Center for Disability Research, 11614 Riyadh, Saudi Arabia. [6]Department of Electronics Engineering, University of Rome "Tor Vergata", 00133 Rome, Italy. [7]Department of Information and Communication Technology, Islamic University, Kushtia 7003, Bangladesh. [8]Machine Learning-aided Wireless Communications (WCML) Research Laboratory, Islamic University, Kushtia 7003, Bangladesh. ✉email: khairul.ice06@gmail.com; patrizia.livreri@unipa.it; alibakhshikenari@ing.uniroma2.it; sipon@ict.iu.ac.bd

Magnetic Resonance Imaging (MRI) has become the gold standard for brain tumor visualization with conventional therapies and rehabilitation because of its superior ability to detail soft tissue structures. Fluid-Attenuated Inversion Recovery (FLAIR) sequences provide enhanced contrast between tumor-related edema and healthy brain tissue by suppressing cerebrospinal fluid (CSF) signals. Despite this advantage, segmenting tumors from FLAIR MRI images remains a challenging task because of the variability in tumor morphology, the presence of artifacts, and the difficulty in distinguishing non-enhancing tumor regions from the surrounding tissue[4].

Deep learning (DL)-based methods, particularly convolutional neural networks (CNNs) such as U-Net[5], have shown remarkable success in medical image segmentation by leveraging an encoder–decoder architecture with skip connections that preserve spatial resolution. However, CNNs often struggle to capture the long-range dependencies necessary for precise tumor segmentation, because they are predominantly designed to model local features[6]. Transformer-based models that utilize self-attention mechanisms have been introduced to address this issue, thereby enabling the capture of global contextual information and long-range dependencies[7,8]. Although these models have demonstrated promising results, they require substantial amounts of training data and lack the inductive biases of CNNs, such as locality and translation invariance, which are important for medical image segmentation[9].

Hybrid architectures combining the strengths of CNNs and transformers, such as TransUNet[10] and Swin-UNet[11], have emerged in recent years to harness both the local feature extraction and global attention capabilities. However, despite their success, these methods still struggle with effective feature extraction across multiple scales, particularly in the context of brain tumor segmentation, where tumor boundaries are often poorly defined owing to the heterogeneity of the tumor and surrounding tissue characteristics.

In this paper, we introduce DSIT-UNet (Dual-Stream Iterative Transformer-based U-Net), a novel architecture specifically designed for brain tumor segmentation from FLAIR MRI images. Our proposed method addresses the limitations of previous models by incorporating a dual-stream encoder that learns features at multiple scales and an iterative attention mechanism that progressively refines the feature representations. Additionally, we introduced a novel Transformed Spatial-Hybrid Attention Optimization (TSHAO) module that effectively balances the need to preserve local spatial details while capturing the global context. This module improves the delineation of tumor boundaries, making it particularly well-suited for the complex and heterogeneous nature of brain tumors. The proposed approach was evaluated on three benchmark datasets, demonstrating superior performance compared with existing state-of-the-art (SOTA) methods, with clear improvements in both segmentation accuracy and robustness.

Although DL has made significant strides in brain tumor segmentation, challenges remain, particularly when dealing with the complexity and heterogeneity of brain tumor shapes and locations. Our study addresses these challenges by introducing an architecture that synergistically combines the strengths of CNNs and transformers, allowing for more accurate and contextually aware tumor delineation. The proposed DSIT-UNet not only advances SOTA brain tumor segmentation, but also provides a framework that can be adapted to other medical imaging tasks that require precise boundary delineation in the presence of complex, heterogeneous structures.

Contribution

We present DSIT-UNet, a Dual-Stream Iterative Transformer UNet that introduces the following key innovations to advance medical-image segmentation:

- **Dual-stream encoder architecture:** We propose a novel encoder that leverages both even and odd kernel convolutions for enhanced multi-scale feature extraction. Even kernels capture continuous structures and global context, whereas odd kernels focus on boundary transitions and fine-grained details. This dual-stream design addresses the fundamental limitation of single-stream architectures, which struggle to balance tumor region homogeneity with precise boundary delineation, which is a critical challenge in brain tumor segmentation.
- **Iterative transformer-based attention:** Unlike conventional single-pass attention mechanisms, we introduce Spatial Cross-Attention (SCA), Channel Cross-Attention (CCA), and Multi-Head Self-Attention (MSA) in an iterative refinement framework. This allows progressive feature enhancement, resulting in richer spatial representations and more precise segmentation. Ablation studies demonstrated a significant Mean IoU improvement, from 83.78% to 89.78%, confirming the effectiveness of our iterative approach.
- **Transformed spatial-hybrid attention optimization (TSHAO):** Our TSHAO module optimally fuses hierarchical features, preserving local details while maintaining global context, a common trade-off in transformer-based segmentation models. This enables accurate tumor boundary delineation while retaining anatomical awareness, leading to a 9.16% improvement in the Mean IoU when integrated with our dual-stream encoder.
- **Enhanced interpretability with saliency maps:** To address the "black-box" nature of DL models in medical imaging, we incorporate saliency map visualization to highlight critical tumor regions. This not only improves model transparency but also aligns DSIT-UNet's predictions with clinically relevant features, such as tumor boundaries and internal heterogeneity patterns, reinforcing trust in clinical deployment.

Collectively, these innovations enable DSIT-UNet to achieve state-of-the-art performance across multiple benchmark datasets (TCIA, BraTS 2019, and BraTS 2020), delivering up to 13.17% higher Mean IoU than existing methods while maintaining a parameter-efficient design (15.97M DSIT-UNet vs. 177.91M TransUNet).

Organization

This paper is structured to provide a comprehensive exploration of our research on lower-grade glioma detection using advanced AI techniques: section "Literature review" presents a critical review of the current literature,

providing a detailed background on glioma detection methods and highlighting the gaps our research aims to address. Section "Proposed methodology" details our methods, including the novel U-shaped architecture incorporating data preprocessing methods, and the proposed model architectures that include attention components. Section "Experiments setup" describes the empirical evaluation of our proposed DSIT-UNet model, including a description of the dataset, evaluation metrics, and implementation details. Section "Results" reports our findings with both quantitative and qualitative analyses of the performance of our model in lower-grade glioma detection. This was followed by a comprehensive discussion that interpreted the results in a broader research context. Section "Discussion" provides an in-depth analysis of the model's architectural contributions, clinical implications, and a comparative assessment with existing approaches. Section "Limitations and future directions" critically examines the limitations of our approach and outlines avenues for future research to address these challenges. Finally, section "Conclusion" summarizes our key contributions and their potential impact on clinical practice in this rapidly evolving field.

Literature review

Recent advances in machine learning have significantly improved the segmentation and classification of brain tumors with conventional therapies and rehabilitation. Cao et al.[12] demonstrated the effectiveness of Support Vector Machine (SVM) models in differentiating Lower-Grade Gliomas (LGG) from glioblastomas (GBM) using MRI-derived tumor location and volume features. Their model achieved impressive results with an AUC above 0.90, and a classification accuracy exceeding 85%, although it showed limitations in handling heterogeneous tumor morphology and imaging artifacts. The advent of DL has led to substantial improvements in the field. Pereira et al.[13] introduced a Convolutional Neural Network (CNN) incorporating kernel-based segmentation for enhanced feature extraction. Although this approach improved upon traditional methods, it remained susceptible to performance degradation in lower-quality MRI scans and exhibited global dependency issues. Hybrid architectures have also been explored to leverage complementary model strengths. Jlassi et al.[14] combined U-Net and SegNet architectures for lower-grade glioma segmentation, capitalizing on U-Net's precise segmentation capabilities and SegNet's computational efficiency. Despite its improved accuracy, this hybrid approach has struggled with global dependency and complex tumor structures. Context-aware approaches have shown promising results in addressing the segmentation challenges. Liu et al.[15] developed the Context-Aware Network (CANet), which integrated graph convolutional networks with context-guided conditional random fields. The model achieved Dice Scores of 0.767, 0.898, and 0.834 for different tumor regions in the BraTS2019 validation, although additional computational iterations were required for optimal performance. Recent attention-based approaches have focused on addressing the intensity-based segmentation challenges. Liu et al.[16] introduced the Attention-Based Multimodal Glioma Segmentation (AMMGS) model, incorporating multi-attention layers within an enhanced 3D U-Net architecture. The model achieved competitive Dice Scores across multiple tumor regions (WT: 0.7803, TC: 0.8831, ET: 0.8172) on BraTS2020, with similar performance on BraTS2019 (0.7675, 0.8925, 0.8110). However, challenges remain in terms of precise tumor boundary delineation and rehabilitation in highly variable cases. Three-dimensional approaches have further advanced this field, as demonstrated by Zhang et al.[17], who developed a 3D U-Net with weighted patch extraction for segmenting tumor subregions. Their model achieved a Dice Score of 0.81 on BraTS2018 validation, although precise boundary delineation remained challenging despite improved class balance through weighted patch extraction.

Despite these advances, current approaches still face significant challenges in achieving robust and accurate brain tumor segmentation, particularly in boundary delineation and handling of complex tumor morphologies with conventional therapies and rehabilitation. Although attention mechanisms and multi-scale feature learning have shown promise, existing methods often struggle to effectively balance local spatial details with global contextual information. Additionally, most current approaches lack an iterative refinement mechanism that can potentially improve segmentation accuracy through progressive feature enhancement. To address these limitations, we propose DSIT-UNet, a novel architecture that combines dual-stream encoding with an iterative attention mechanism. Our approach introduces the TSHAO module, which specifically targets the challenge of maintaining detailed spatial information, while incorporating broader contextual features. This architecture advances the field by offering a more sophisticated approach for feature refinement and boundary delineation, which is particularly crucial for the heterogeneous nature of brain tumors on FLAIR MRI with conventional therapies and rehabilitation.

Proposed methodology

The workflow of the proposed methodology is shown in Fig. 1. Our approach consists of two primary components: data processing and innovative model architecture. During the preprocessing phase, we ensured the optimal model performance by correctly formatting and normalizing the input data. Our architecture integrates several sophisticated components, with transformer encoder blocks serving as its foundation. This base architecture is complemented by a DSIT encoder, which enables parallel processing of distinct information streams. We employed multiscale feature extraction through a convolutional patch embedding mechanism that operates across multiple encoder stages, effectively capturing both fine- and coarse-grained patterns in the data. The architecture's cornerstone is a novel Iterative Transformer (IT) block that progressively refines feature representations through multiple attention passes. IT works synergistically with two specialized transformer components: the Cross-Attention (CA) block, which facilitates interaction between different feature spaces, and the Latent Transformer (LT) block. The LT comprises regular transformer blocks that process information within the individual feature spaces. The decoder module synthesizes processed information to produce the final output.

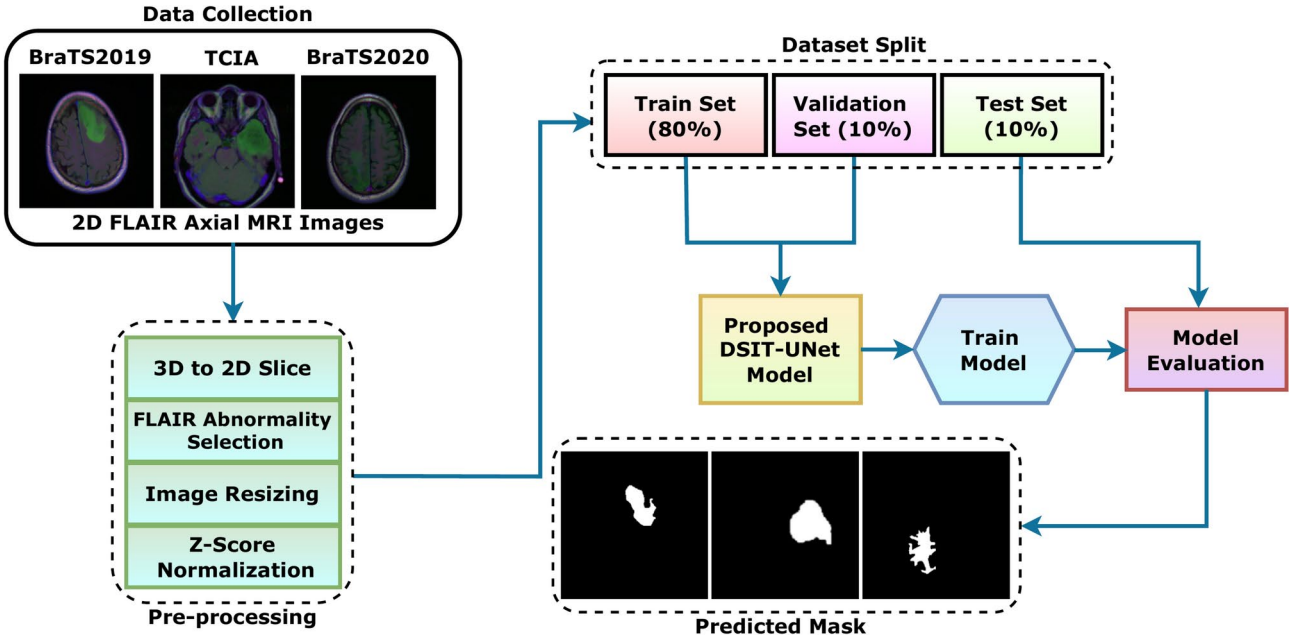

Fig. 1. DSIT-UNet workflow: End-to-end brain tumor segmentation pipeline utilizing dual-branch encoding and iterative attention mechanisms for processing 2D FLAIR axial MRI slices, validated across BraTS2019, TCIA, and BraTS2020 datasets.

Data preprocessing

To ensure optimal model performance and consistency across the dataset, we implemented a comprehensive preprocessing pipeline comprising of four essential stages: 2D slice extraction, FLAIR abnormality selection, image resizing, and Z-score normalization.

2D slice with FLAIR abnormality selection

Given that our DSIT-UNet operates as a 2D network, we processed each 3D medical image slice-by-slice, specifically selecting axial slices from available axial, coronal, and sagittal 3D images. We implemented a selective approach focusing on images containing FLAIR abnormalities in which only slices with corresponding segmentation masks showing FLAIR abnormalities were retained in our dataset. This selection process ensured that the model concentrated on the relevant features for abnormality identification, while images with entirely black masks (indicating no abnormalities) were excluded. This approach maintained the focus of our training data on clinically significant cases.

Image resizing

To ensure a uniform input to our model, all images were standardized to consistent dimensions of 256×256 pixels across all three modalities. This standardization is crucial for maintaining spatial consistency across datasets and optimizing computational efficiency, while preserving essential structural information.

Z-score normalization

We applied Z-score normalization to all images to address the intensity variations arising from magnetic field non-uniformity across different scanners. This process involves subtracting the mean from each voxel and dividing it by the standard deviation, thereby achieving zero mean and unit variance for each brain image. This normalization step is crucial for addressing scanner-dependent intensity variations, improving model training efficiency, enhancing the model's generalization capabilities, and facilitating faster convergence during training.

Notably, the BraTS2020 dataset underwent initial preprocessing by competition organizers, including isotropic resolution achievement, general spatial alignment, skull removal, and co-registration. These preliminary steps provide a standardized foundation for the subsequent preprocessing pipeline.

Proposed model

Dual-stream iterative transformer (DSIT) encoder

The overall structure of the model adopted a U-shaped architecture[5]. A single encoder comprises two primary blocks, a multiscale patch extractor (MSPE) block and an Iterative Transformer (IT) block, as illustrated in Fig. 2. The MSPE blocks extract spatiotemporal features from input images using n spatial convolutional layers. Subsequently, IT blocks employ causal attention and emphasize crucial temporal information through an MSA mechanism.

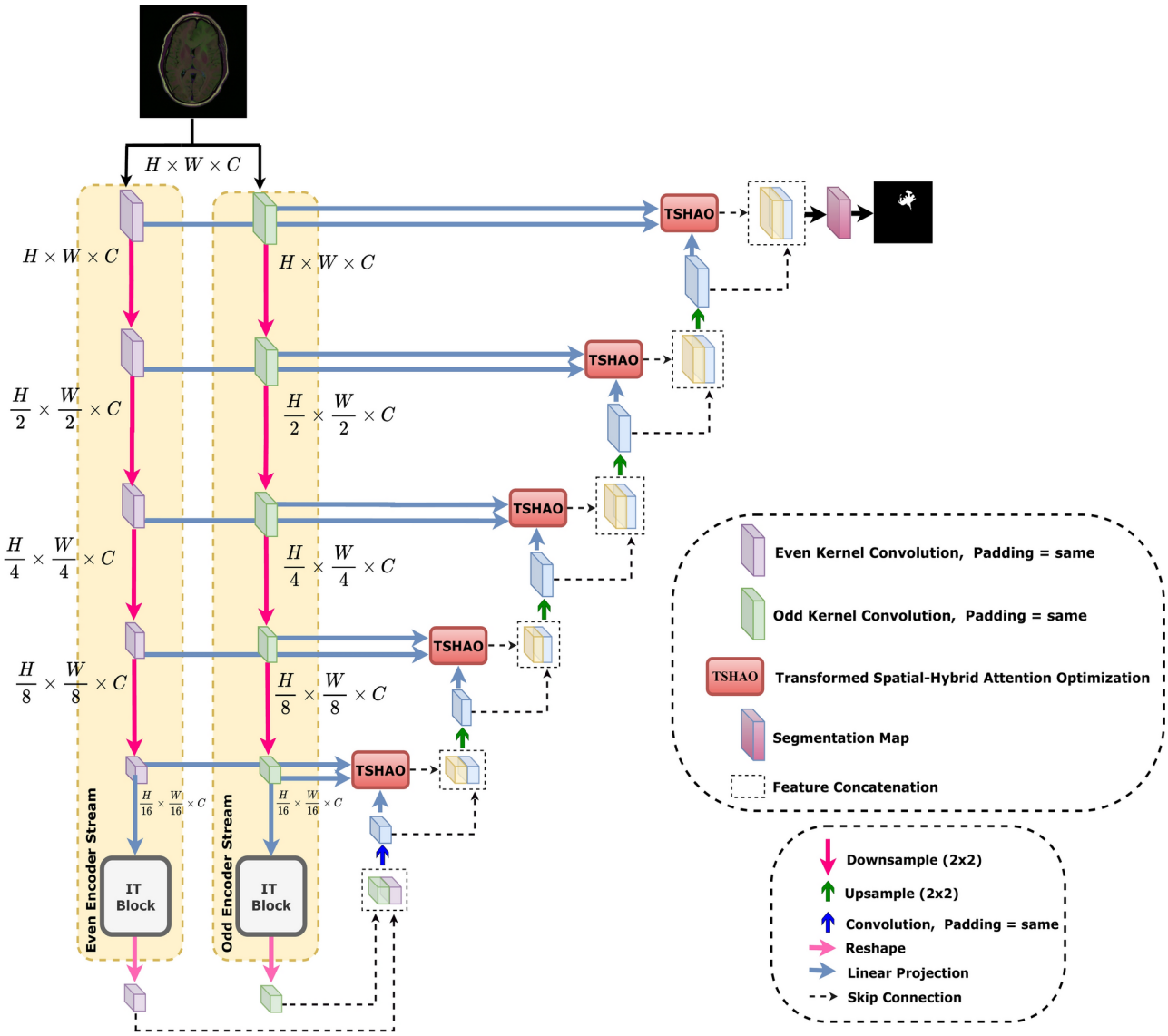

Fig. 2. DSIT-UNet architecture: Dual-stream encoder utilizing even-odd kernels, iterative attention mechanisms (SCA, CCA, MSA), and multi-scale feature extraction pathways for tumor segmentation.

Multi-scale patch extractor (MSPE) block

MSPE utilizes convolution blocks to extract and process localized segments of the input images as patches, which are then fed into the IT block. The MSPE block in each encoder employs a unique approach: one encoder uses odd kernel sizes in its convolutional layers, whereas the other uses even kernel sizes. Odd-sized kernels, such as 3×3 kernels, provide a symmetric receptive field that avoids pixel shifts and maintains spatial alignment across the convolutional layers. In contrast, even-sized kernels, such as 2×2 and 4×4 kernels, can reduce the computational complexity and memory consumption while extending receptive fields, especially when combined with symmetric padding to mitigate shift issues[18]. This differential-kernel strategy enables diverse feature extractions at varying scales. The 2D convolutional layer used ReLU activation, layer normalization, and dropout. The multiscale patch extraction process involved 2D max pooling. Finally, the patches were reshaped to $(\frac{H}{2^{i-1}} \times \frac{W}{2^{i-1}}, C_i)$ before normalization and input into the IT block.

$$T_i = \mathrm{LN}(R(P(\sigma\{\mathrm{Conv2D}(E_i)\}))) \tag{1}$$

$R(\cdot)$ and $P(\cdot)$ denote reshaping and pooling operations, respectively. Tokens $T_i \in \mathbb{R}^{P_i \times C_i}$ represent the flattened patches for the i-th encoder stage E_i, where P_i is the flattened patch dimension $(\frac{H}{2^{i-1}} \times \frac{W}{2^{i-1}})$ common across all T_i and C_i denotes the number of patches. Convolution captures the spatial information during patch extraction, eliminating the need for positional information of the extracted patches.

Iterative transformer (IT) block

The IT block architecture integrates a CA block with an LT block, as illustrated in Fig. 3. The design process begins with the MSPE block, which extracts localized patches that are subsequently processed by the IT block. Within this structure, LT captures long-range dependencies to refine the feature representations, whereas the CA block facilitates the integration of multiscale features. This architectural synergy enhances the segmentation precision by effectively balancing fine-grained details with global contextual coherence.

Inspired by[9], we implemented an IT block, owing to its architectural efficiency and scalability. The Cross-Attention block employs a query (Q) network that operates on latent-space input, whereas the key (K) and value (V) networks process the direct input. This configuration enabled the attention mechanism to effectively condense high-dimensional inputs into compressed latent representations. The subsequent Latent Transformer block functions as a standard transformer encoder, incorporating MSA, a Feed-Forward Network (FFN), and residual connections, followed by cross-attention blocks.

The operational mechanism of the architecture parallels that of a Recurrent Neural Network (RNN)[19], with the key distinction being its depth-wise rather than temporal unrolling when processing input. The architecture implements a cross-attention input projection, employs bottleneck latent dimensionality, and uses a recurrent latent transformer core. Weight sharing in this context achieves effects analogous to those of traditional transformers, while the extensive receptive field enables robust feature extraction, thus enhancing segmentation performance without compromising computational efficiency[9].

Cross-attention

The dual-branch encoder architecture implements two parallel attention mechanisms: SCA and CCA. This dual-attention approach provides complementary analytical perspectives: channel attention identifies salient features, whereas spatial attention localizes their positions, resulting in comprehensive image understanding. In the cross-attention mechanism, queries are derived from the latent array, whereas the key and value vectors are generated from the encoded image. The SCA is formally defined as

$$\mathrm{SCA}(\mathrm{Q},\mathrm{K},\mathrm{V}) = \mathrm{Softmax}\left(\frac{\mathrm{QK}^T}{\sqrt{d_k}}\right)\mathrm{V} \tag{2}$$

where $Q \in \mathbb{R}^{1\times L\times C_i}$, $K \in \mathbb{R}^{N\times P\times C_i}$, and $V \in \mathbb{R}^{N\times P\times C_i}$ represent the projected query, key, and value, respectively. L denotes the latent array dimension, a predetermined hyperparameter. The scaling factor was defined as $\frac{1}{\sqrt{d_k}} = \frac{1}{\sqrt{C_i}}$.

For Channel Cross-Attention (CCA), the computation involves permuting the tokens of the secondary encoder T_l and applying attention across the channel dimension:

$$\mathrm{CCA}(\mathrm{Q}_i,\mathrm{K}_i,\mathrm{V}_i) = \mathrm{Softmax}\left(\frac{\mathrm{Q}_i\mathrm{K}_i^T}{\sqrt{d_k}}\right)\mathrm{V}_i \tag{3}$$

where $K_i \in \mathbb{R}^{N\times C_i\times P}$ and $V_i \in \mathbb{R}^{N\times C_i\times P}$ denote projected keys and values, respectively.

Self-attention

The Self-Attention (SA) block implements the MSA mechanism that processes normalized input features by utilizing the CA output for both Q and K. Following the standard transformer architecture[20], the block executed sequential attention operations. The Spatial Self-Attention (SSA) computations were formulated as follows:

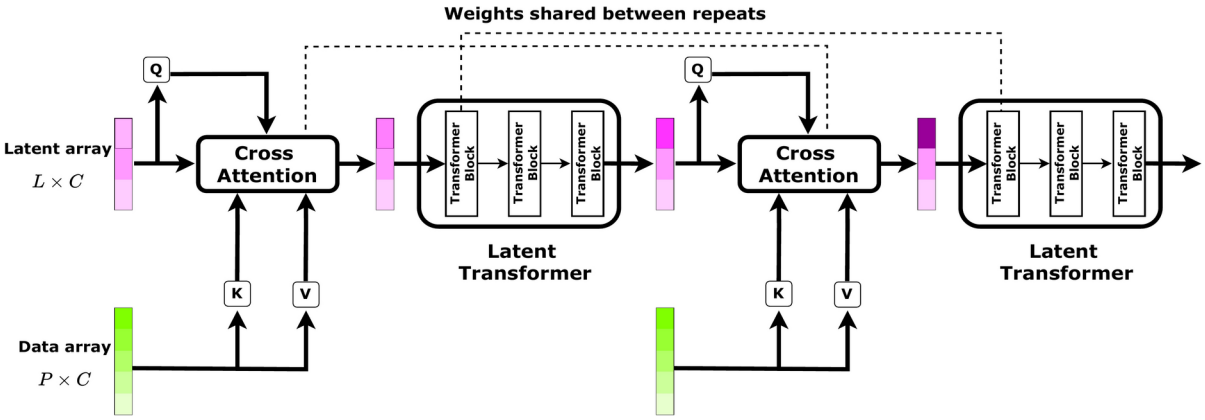

Fig. 3. Iterative transformer architecture: cross-attention patch mapping with weight-shared self-attention refinement.

$$\begin{aligned} \mathrm{SSA}_j^{(i)} &= \mathrm{MSA}(\mathrm{LN}(\mathrm{SCA}_j^{(i)})) + \mathrm{SCA}_j^{(i)} \\ \mathrm{SSA}_j'^{(i)} &= \mathrm{FFN}(\mathrm{LN}(\mathrm{SSA}_j^{(i)})) + \mathrm{SSA}_j^{(i)} \end{aligned} \tag{4}$$

$$\begin{aligned} \mathrm{SSA}_j^{(l)} &= \mathrm{MSA}(\mathrm{LN}(\mathrm{CCA}_j^{(l)})) + \mathrm{CCA}_j^{(l)} \\ \mathrm{SSA}_j'^{(l)} &= \mathrm{FFN}(\mathrm{LN}(\mathrm{SSA}_j^{(l)})) + \mathrm{SSA}_j^{(l)} \end{aligned} \tag{5}$$

Here, SCA, CCA, and SSA represent real-valued matrices, with dimensions $(N \times L \times C_i)$ corresponding to the outputs from the CA and MSA blocks. The block indices are denoted by the superscripts $i = 1, 2, \ldots n$ and $l = 1, 2, \ldots n$, whereas layer-specific operations are indexed by $j = 1, 2, \ldots n$. The process iterates through a fixed number of transformer blocks ($n = 3$), with $\mathrm{SSA}_j'^{(i)}$ and $\mathrm{SSA}_j'^{(l)}$ updated recursively during each iteration.

The complete IT block mechanism can be expressed concisely as

$$\begin{aligned} L_i &= f_T(f_{CA}(L_{i-1}, D)), \quad i \in \{1, 2, \ldots, n\} \\ L_n &= \underbrace{f_T \circ f_{CA} \circ \cdots \circ f_T \circ f_{CA}}_{n \text{ iterations}}(L_0, D) \end{aligned} \tag{6}$$

This final equation describes the forward pass of the IT block, where the latent array undergoes iterative updates through multiple applications of CA and transformer functions. The initial latent array L_0 is updated at each step, whereas the encoded patches D (data array) facilitate latent array interaction through the cross-attention mechanism. This process culminates in the final latent array L_n after n iterations. The composition operator $\circ$ indicates a sequential function application beginning with cross-attention (f_{CA}), followed by the transformer (f_T) and repeated n times.

Decoder

As illustrated in Fig. 2, the initial decoder input is formed by reshaping and merging the outputs of the IT blocks. In each subsequent decoder stage, the input features undergo upsampling by a factor of 2 using Conv2DTranspose and ReLU activation. These upsampled features were concatenated with the corresponding skip-connection feature maps of the encoder. The combined features were processed using the TSHAO module within the skip connection.

This design has been adopted for several reasons, **I**) It efficiently utilizes encoder features through strategically placed skip connections at each stage. **II**) The TSHAO module in the skip connections facilitates the development of long-range dependencies and enhances global context interaction within the decoder. **III**) The progressive upsampling and feature refinement process improves decoding performance. Section V-B thoroughly explores and discusses the significant impact of incorporating the TSHAO module into the skip connections.

Transformed spatial-hybrid attention optimization (TSHAO) module

The critical challenge in multiscale feature representation learning is the effective fusion of the output features from our dual-branch encoder. Although a simplistic approach might involve feature concatenation followed by convolution, such a method fails to capture long-range dependencies and global context connections across different scales. To address this limitation, we introduced a novel TSHAO module, as illustrated in Fig. 4. The TSHAO module leverages the MSA mechanism to facilitate efficient and effective interactions between the multiscale features.

In our TSHAO module architecture, Q and K were derived from dual encoder branches. V is mapped from the upsampled features of the preceding decoder stage. This configuration enhances the multiscale feature representation and promotes cross-scale interactions through the attention mechanism. While conventional

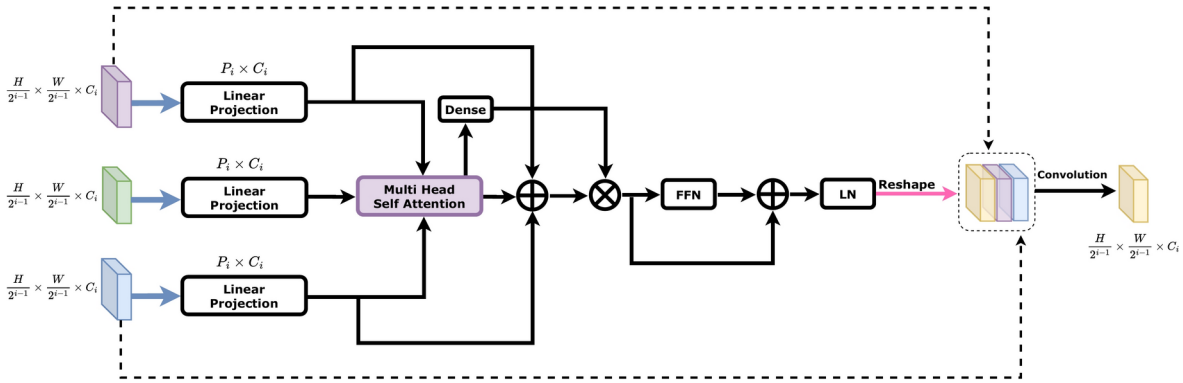

Fig. 4. Transformed spatial-hybrid attention optimization (TSHAO) block: Multi-scale feature processing utilizing depthwise convolutions for *Q-K* pairs from dual-stream encoders and *V* from decoder, with dynamic attention weighting for comprehensive feature dependency capture.

self-attention typically employs linear projections, recent research has successfully integrated convolutions into self-attention mechanisms, introducing locality and reducing computational complexity[21,22]. In particular, depth-wise convolutions have gained prominence in self-attention architectures, owing to their ability to capture local information with minimal computational overhead[23–26]. Inspired by these advancements, we replaced all linear projections in the TSHAO module with depthwise convolutional projections:

$$\begin{aligned} \mathrm{Q}_i &= \mathrm{DConv1D}_Q(T_i) \\ \mathrm{K}_i &= \mathrm{DConv1D}_K(T_l) \\ \mathrm{V}_i &= \mathrm{DConv1D}_V(D_i) \end{aligned} \tag{7}$$

Where $Q \in \mathbb{R}^{P_i \times C_i}$, $K \in \mathbb{R}^{P_i \times C_i}$ and $V \in \mathbb{R}^{D_i \times C_i}$ denote the projected query, key, and value, respectively. T_i and T_l denote the flattened features from the encoder branches, and D_i represents the upsampled and reshaped features with the same dimensions as those in the preceding decoder stage. The attention mechanism is then applied as follows:

$$\begin{aligned} z_l &= \mathrm{MSA}(\mathrm{Q}_i, \mathrm{K}_i, \mathrm{V}_i) + \mathrm{Q}_i + \mathrm{K}i \\ z_{l+1} &= z_l \odot \mathrm{Dense}(\mathrm{MSA}(\mathrm{Q}_i, \mathrm{K}i, \mathrm{V}i)) \\ z_{l+2} &= \mathrm{LN}(\mathrm{FFN}(z_{l+1}) + z_l) \end{aligned} \tag{8}$$

where the dynamic MSA output $z_{l+2} \in \mathbb{R}^{P_i \times C_i}$, Then, we reshaped z_{l+2} to dimensions $(\frac{H}{2^{i-1}}, \frac{W}{2^{i-1}}, C_i)$, where H and W are the original height and width of the input, respectively. This reshaped tensor was concatenated with the output of the dual encoder (E_i and E_l) from the same stage. Finally, we applied 2D convolution, followed by ReLU activation, to these concatenated features.

$$z_{out} = \mathrm{ReLU}(\mathrm{Conv2D}(\mathrm{concat}(R(z_{l+2}), E_i, E_l))) \tag{9}$$

Where $R(\cdot)$ denotes the reshaping operation, $\mathrm{concat}(\cdot)$ represents the concatenation along the channel dimension, $\mathrm{Conv2D}(\cdot)$ is a 2D convolution operation, and $\mathrm{ReLU}(\cdot)$ is the Rectified Linear Unit activation function.

Experiments setup

We conducted a comprehensive evaluation of three different medical image segmentation tasks to evaluate the effectiveness and adaptability of the proposed DSIT-UNet. Our experiments used several widely recognized and publicly accessible datasets, enabling fair comparison with other SOTA methods in the field. This section provides an overview as **Dataset descriptions**. Summary of medical image datasets used for segmentation experiments. **Evaluation metrics**. Performance measures for model evaluation and comparison. **Implementation details**. Hardware, Software, and Key Experimental Parameters.

Dataset descriptions

This study validated our model using three distinct datasets: the Multimodal Brain Tumor Segmentation Challenge (BraTS) 2019 and 2020 datasets[27,28] and data from The Cancer Imaging Archive (TCIA)[29]. These datasets were selected to ensure comprehensive validation across various glioma types, imaging modalities, and clinical contexts, thereby comprehensively evaluating the generalizability and clinical applicability of the model.

BraTS 2019 and 2020 datasets

The BraTS datasets[27,28,30] specifically target brain tumor segmentation, featuring multimodal MRI scans of glioma patients (Fig. 5). The BraTS2019 dataset comprises a training set of 335 cases and a validation set of 125 cases, whereas BraTS2020 includes 369 training cases, 125 validation cases, and 166 test cases. These datasets are invaluable for testing models in both high- and low-grade glioma cases, offering a mix of tumor grades and locations, which are common challenges in brain tumor segmentation. The imaging data included

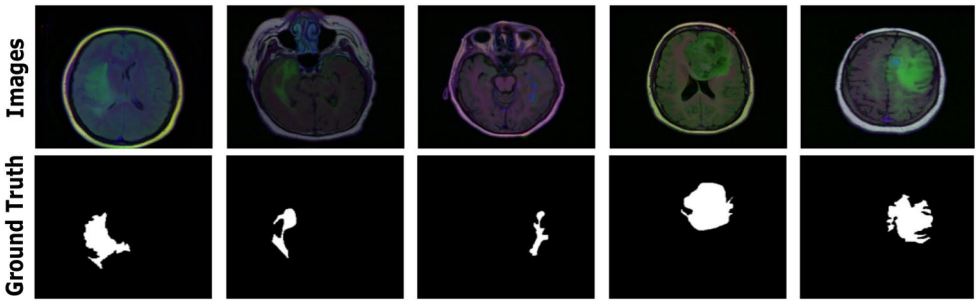

Fig. 5. Sample images from the BraTS2020 dataset showing brain MRI axial scans and corresponding ground truths. The diverse and complex tumor morphologies present challenges for DL models in accurate feature extraction and tumor delineation.

T1-weighted, T1c (contrast-enhanced), T2-weighted, and FLAIR sequences, which provided complementary information regarding the anatomical and pathological features of the tumor. The BraTS datasets also include expert-annotated tumor subregions, such as the enhancing core, non-enhancing core, and edema, which allow for a detailed evaluation of the model's performance in segmenting these key tumor components. The diversity of these tumor characteristics and availability of well-annotated data were key to selecting BraTS, ensuring that the model could be assessed across a wide range of gliomas with different levels of malignancy and anatomical presentations.

TCIA dataset

The inclusion of the TCIA dataset[29,31], which consists of MRI scans from 110 patients with lower-grade gliomas (LGGs, WHO grades II and III) from The Cancer Genome Atlas (TCGA), further broadens the scope of the model's validation. The TCIA dataset adds complexity by integrating genomic data from TCIA, which allows radiogenomic studies to explore the relationship between tumor imaging features and underlying genetic alterations. This is particularly useful for glioma subtyping, where genomic features such as IDH mutation status and 1p/19q co-deletion play a crucial role in treatment planning and prognosis. The TCIA dataset includes preoperative MRI scans (T1-weighted, T1c, T2-weighted, and FLAIR) with expert segmentation labels for tumor regions such as the enhancing tumor core, non-enhancing tumor core, and peritumoral edema. By incorporating both imaging and genomic data, TCIA enables a comprehensive assessment of the model's ability to predict patient outcomes and genomic subtypes, which is an important step towards personalized medicine.

In all three datasets, the focus was on FLAIR modality images, which are particularly useful for identifying peritumoral edema, a crucial feature in brain tumor assessment. The datasets were chosen not only for their quality and availability of expert annotations, but also for their diversity in glioma cases, ranging from aggressive high-grade gliomas to slower-growing lower-grade gliomas, with a focus on providing a well-rounded evaluation of the model's performance across various clinical contexts.

Evaluation metrics

To evaluate our proposed DSIT-UNet, we employed comprehensive quantitative metrics commonly used in medical image segmentation. The evaluation framework incorporates the following standard metrics, calculated using true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN):

$$\begin{aligned} \text{MeanIoU} &= \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i}{TP_i + FP_i + FN_i} \\ \text{Precision} &= \frac{TP}{TP + FP} \\ \text{Recall} &= \frac{TP}{TP + FN} \\ \text{mDice} &= 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \\ \text{Accuracy} &= \frac{TP + TN}{TP + TN + FP + FN} \end{aligned} \tag{10}$$

Implementation details

To improve model performance, we used a multiscale training strategy instead of traditional data augmentation. The model was trained using the binary cross-entropy loss as the objective function. Unlike approaches that utilize deep supervision[32], we rely solely on the final loss for training. The best hyperparameters for the proposed DSIT-UNet model is given in Table 1. This section provides a detailed overview of the architecture of the proposed model. All experiments were performed on NVIDIA Tesla P100 GPUs with 16GB of memory, using the TensorFlow framework.

Results

Quantitative performance analysis

We evaluated the DSIT-UNet using three benchmark datasets: TCIA, BraTS 2019, and BraTS 2020. On the TCIA dataset, DSIT-UNet achieved exceptional performance with a Mean IoU of 0.9521, precision of 0.9591, recall of 0.9655, and mDice of 0.9623 (Table 2). These results demonstrate substantial improvements over previous SOTA models, with performance gains of 7.51%, 9.77%, and 13.17% compared to DoubleU-Net[33], TransUNet[10], and DS-Trans UNet[34], respectively. While DS-Trans UNet achieved marginally higher precision, DSIT-UNet demonstrated superior performance across all evaluation metrics.

The effectiveness of the model was further validated using BraTS datasets. For BraTS 2019, DSIT-UNet achieved a mean IoU of 96.08%, precision of 95.34%, recall of 95.55%, and mDice of 95.44%. In BraTS 2020, the model significantly outperformed the existing SOTA method (DS-Trans UNet), with improvements of 10.20%, 10.98%, and 15.0% in the mean IoU, precision, and recall, respectively. These consistent improvements across diverse datasets underscore the robustness and generalizability of the proposed model for clinical applications.

Beyond the segmentation accuracy, we evaluated the computational efficiency and complexity of DSIT-UNet against state-of-the-art models (Table 3). DSIT-UNet achieved superior performance while maintaining a balanced trade-off between efficiency and complexity. It requires significantly fewer parameters (15.97M) than transformer-based models, such as TransUNet and DS-Trans UNet, with a shorter inference time (24 vs. 29 s).

Hyperparameter	Odd encoder stream	Even encoder stream
MSPE Block		
Image Size	256 × 256 × 3	256 × 256 × 3
Kernel Size	3 × 3	2 × 2
No. of Conv Filters	256	256
No. of Conv Layers	5	5
Padding	Symmetric	Symmetric
Conv Layer Activation	ReLU	ReLU
IT Block		
Patch Size	16 × 16	16 × 16
Data Dimension	256	256
Latent Dimension	256	256
Transformer Block	3	3
FFN Dimension	256	256
No. of Skip Connections	7	7
FFN Activation	GELU	GELU

TSHAO module		Training parameters	
Hyperparameter	**Value**	**Hyperparameter**	**Value**
Projection Dimension	256	Epochs	70
Number of Heads	2	Learning Rate	0.00014
FFN Activation	GELU	Normalization	LayerNorm

Table 1. Hyperparameter configuration and architectural specifications of the proposed model implementation.

Although its FLOP count (385.05G) is higher than that of the conventional U-Net variants, the performance gains justify the computational cost.

To ensure a fair comparison, we independently reproduced the SOTA models under identical experimental conditions, including dataset splits, preprocessing steps, and evaluation metrics.

Qualitative performance analysis

Qualitative analysis (Fig. 6) provides visual evidence of DSIT-UNet's superior segmentation capabilities. The model demonstrated exceptional performance in two critical aspects: accurate tumor localization and precise boundary delineation between tumor and healthy tissues. These visual results complement the quantitative metrics and validate the practical applicability of the model in clinical settings.

Feature visualization

To provide deeper insight into the learning mechanisms of DSIT-UNet, we analyzed the representations of the features of two crucial architectural components. First, the odd and even kernel features of the MSPE block (Fig. 7) demonstrate the complementary nature of our dual-stream approach. The even kernel features excel at capturing continuous structural patterns and broader contextual information, whereas the odd kernel features specialize in detecting boundary transitions and fine-grained details. This comprehensive feature extraction strategy enables even kernels to focus on regional consistency and tissue homogeneity, whereas odd kernels emphasize edge information and local variations in tumor boundaries. Visualization revealed how this complementary approach enhanced the ability of the model to delineate tumor regions from the surrounding healthy tissue to support rehabilitation.

Second, the decoder refines the features (Fig. 8) illustrate the efficacy of our iterative attention mechanism and the TSHAO module. Visualizations show progressive feature refinement through multiple attention mechanisms in which the initial coarse representations evolve into more precise and contextually rich features. The refined features demonstrated enhanced spatial coherence and stronger activation patterns in tumor-relevant regions, validating our architectural choice of incorporating iterative attention and the TSHAO module. This progressive improvement in the feature quality and specificity through successive refinement steps is evident in the visualization results.

Saliency maps as explanable artificial intelligence

To enhance the interpretability and transferability of DSIT-UNet and provide deeper insights into its decision-making process, we employed gradient-based saliency maps as Explanable Artificial Intelligence (XAI)[38]. These maps highlight image regions that significantly influence the segmentation decisions of the model by computing the output gradient of the input image. The generated saliency maps revealed that DSIT-UNet primarily focused on tumor boundaries and internal texture patterns, demonstrating its ability to identify clinically relevant features. An analysis of the saliency maps shown in Fig. 9 showed strong activation patterns along the tumor margins, indicating that the model effectively utilizes edge information for precise boundary delineation. Additionally, the maps exhibited gradual intensity variations within the tumor regions, suggesting that the

Method	TCIA				BraTS2019				BraTS2020			
	Mean IoU (%)	Precision (%)	Recall (%)	mDice (%)	Mean IoU (%)	Precision (%)	Recall (%)	mDice (%)	Mean IoU (%)	Precision (%)	Recall (%)	mDice (%)
U-Net[5]	81.71	89.17	67.60	72.93	81.68	94.10	70.60	80.67	81.73	88.20	67.66	76.58
Res-UNet[35]	85.78	87.70	83.95	85.78	86.01	88.79	84.45	86.57	85.88	87.77	80.95	84.22
Res-UNet++[36]	84.58	87.52	80.58	83.91	80.81	87.58	82.51	84.97	82.08	87.01	80.08	83.40
DeepLabv3+[37]	70.01	80.23	76.68	78.41	73.03	85.11	79.03	81.96	75.05	79.33	80.01	79.67
Double-UNet[33]	85.30	82.02	83.11	82.56	86.47	83.03	85.23	84.12	86.71	80.78	83.24	81.99
TransUNet[10]	87.58	85.36	86.58	85.97	88.08	86.51	87.52	87.01	87.44	86.78	88.08	87.40
DS-TransUNet[34]	87.70	**96.02**	86.78	91.17	87.81	85.08	86.18	85.62	85.68	85.23	84.98	85.11
DSIT-Unet (Ours)	**95.21**	95.91	**96.55**	**96.23**	**96.08**	**95.34**	**95.55**	**96.44**	**95.88**	**96.21**	**96.44**	**96.32**

Table 2. Quantitative evaluation of brain tumor segmentation methods: Comparing DSIT-UNet with SOTA models across TCIA, BraTS2019, and BraTS2020 datasets.

Model	Inference Time (s)	FLOPs (G)	Param (M)
U-Net	20	144.11	12.15
Res-UNet	22	156.01	14.19
Res-UNet++	22	161.45	14.33
DeepLabV3+	19	145.25	10.56
Double-UNet	25	178.07	25.03
TransUNet	31	308.23	177.91
DS-TransUNet	29	358.02	107.56
DSIT-UNet (Ours)	24	385.05	**15.97**

Table 3. Performance and Complexity Comparison of SOTA Segmentation Models.

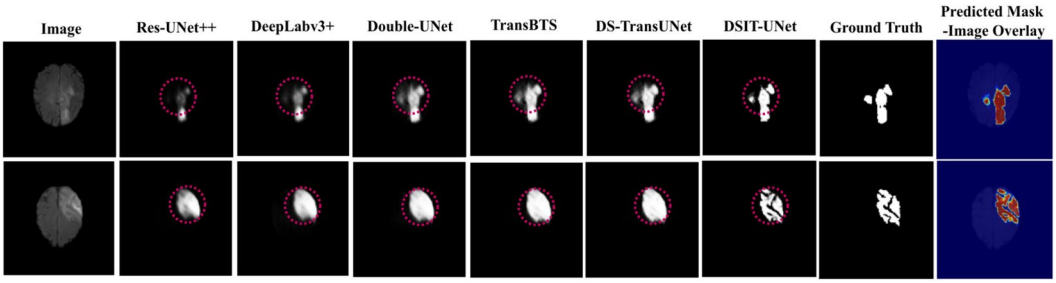

Fig. 6. Qualitative assessment of DSIT-UNet segmentation performance versus SOTA models.

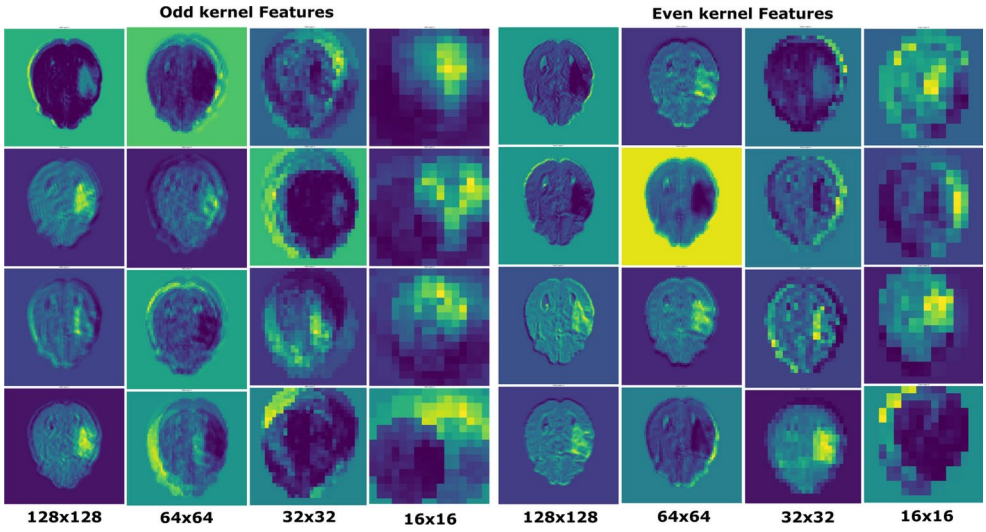

Fig. 7. Extracted Odd and even kernel features by the final four layers of the MSPE, showing diverse learned representations.

model considers internal tissue heterogeneity when making segmentation decisions. This hierarchical attention to both boundary and internal features aligns with clinical expertise, in which radiologists consider both tumor margins and internal characteristics for accurate diagnosis.

Saliency visualization also reveals the robustness of the model to imaging artifacts and normal anatomical structures, as evidenced by minimal activation in non-tumor regions. This selective attention mechanism helps explain the high-precision scores of DSIT-UNet across different datasets. Furthermore, the consistency of saliency patterns across various test cases demonstrates the stable learning of the relevant features of the model, contributing to its generalization capabilities.

Ablation analysis

To validate the effectiveness of our proposed components, we conducted comprehensive ablation studies using the BraTS 2019 dataset.

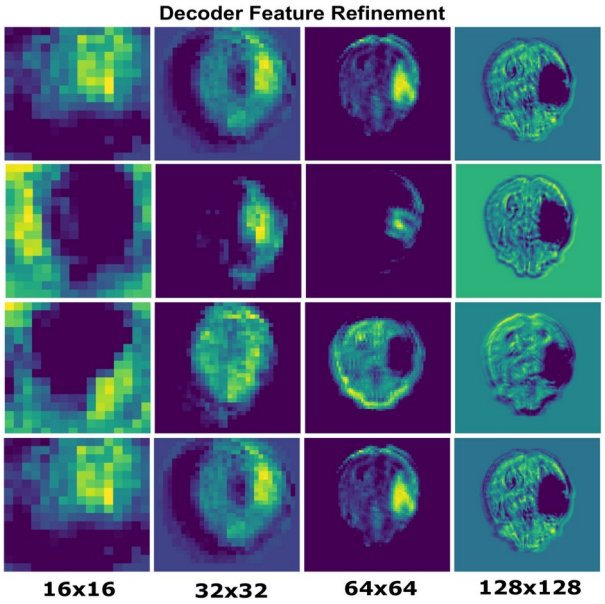

Fig. 8. Visualization of decoder feature refinement, highlighting the enhanced segmentation of tumor boundaries.

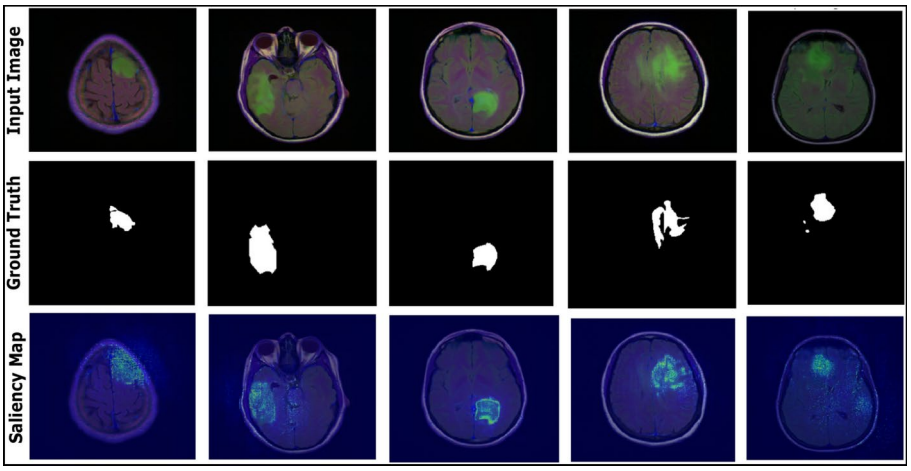

Fig. 9. Visualizing model focus: saliency maps for DSIT-UNet.

Effect of dual encoder, TSHAO, and IT

The results (Table 4) highlight the systematic performance improvements achieved by integrating the key architectural enhancements. The baseline single-encoder model achieved a Mean IoU of 75.01%, thereby demonstrating its fundamental segmentation capability. Incorporating TSHAO into the Single Encoder improved the performance to 78.07%, whereas the addition of IT further elevated the Mean IoU to 83.78%, highlighting the impact of iterative attention. The dual-encoder architecture provided a significant boost (81.62%) with additional gains when combined with TSHAO (87.23%) and IT (89.78%). The full DSIT-UNet model, which integrates both mechanisms, achieved remarkable mean IoUs of 96.08 % and 95.86% mDice, surpassing all prior configurations and validating the synergistic effect of these enhancements.

Effect of kernel configurations

Table 5 presents a comparison of different kernel configurations. The even-kernel setup resulted in a lower recall (70.20%), thereby limiting segmentation completeness. The odd kernel configuration improved the Mean IoU (90.19%) and recall (90.58%). However, DSIT-UNet achieved the best performance with a mixed-kernel configuration, reaching a 95.23% Mean IoU, 96.88% recall, and 96.36% mDice. This result confirms that combining different kernel types optimizes the feature extraction and spatial representation.

Model variants	Param (M)	FLOPs (G)	Accuracy (%)	Mean IoU (%)	mDice (%)
Single Encoder	10.14	155.3	92.03	75.01	76.71
Single Encoder + TSHAO	13.11	160.5	95.41	78.07	80.09
Single Encoder + IT	14.24	175.01	97.01	83.78	85.67
Dual Encoder	13.61	159.8	98.02	81.62	85.03
Dual Encoder + TSHAO	14.67	171.01	98.97	87.23	90.05
Dual Encoder + IT	15.78	205.93	99.31	89.78	92.25
DSIT (Ours)	**15.97**	**385.05**	**99.92**	**96.08**	**95.86**

Table 4. Performance comparison of different model variants.

Kernel type	Kernel size	Dilution factor	Accuracy (%)	Mean IoU (%)	Precision (%)	Recall (%)	mDice (%)
Even	2, 2, 2, 2, 4	2, 2, 2, 2, 2	98.61	88.52	90.00	70.20	78.87
Odd	3, 3, 3, 3, 1	3, 3, 3, 3, 3	98.71	90.19	87.02	90.58	88.76
Mixed	**(2, 2), (3, 3)**	**1, 1, 1, 1, 1**	**99.92**	**95.23**	**95.85**	**96.88**	**96.36**

Table 5. Comparison of DSIT-UNet Performance with Different Kernel Configurations.

Experiment no.	Patch size	Latent dimension	Param # (M)	Encoder–decoder depth	Accuracy (%)	Mean IoU (%)	Precision (%)	Recall (%)	mDice (%)
1	128 × 128	256 × 256	12.89	2–2	97.47	79.84	86.22	75.68	80.61
2	64 × 64	256 × 256	12.96	3–3	98.36	83.65	86.76	79.14	82.78
3	32 × 32	256 × 256	13.71	4–4	98.79	84.45	87.55	86.68	87.11
4	**16 × 16**	**256 × 256**	**15.97**	**5–5**	**99.93**	**95.22**	**95.85**	**96.89**	**96.37**
5	8 × 8	256 × 256	22.61	6-6	98.98	87.15	89.51	89.21	89.36

Table 6. Impact of patch size and encoder–decoder depth on DSIT-UNet performance.

Impact of patch size and encoder–decoder depth

As shown in Table 6, reducing the patch size significantly improved the segmentation performance. A patch size of 16 × 16 with a 5–5 encoder–decoder depth yielded the best results, achieving a Mean IoU of 95.22 % and an mDice of 96.37 %. Increasing the depth beyond this point (6-6) led to a performance decline, likely owing to overfitting and excessive parameter complexity. These findings indicate that the selected patch size and depth effectively balance feature extraction and computational efficiency.

Overall, our ablation study demonstrates that the architectural enhancements of DSIT-UNet, including the dual encoder, iterative attention mechanisms, mixed kernel configurations, and optimized patch size, contribute significantly to its superior segmentation performance.

Discussion

The experimental results demonstrated that DSIT-UNet significantly advanced the SOTA brain tumor segmentation across multiple benchmark datasets. This section explores the architectural contributions, clinical implications, and comparative analysis with existing methods, limitations, and future research directions.

Architectural innovations and their impact

The superior performance of DSIT-UNet is mainly attributed to its novel dual-stream architecture with mixed kernel configurations, which improves feature extraction capabilities. As confirmed by ablation studies, even kernels effectively capture continuous structural patterns and regional consistency, whereas odd kernels excel at detecting boundary transitions and fine-grained details. The complementary feature-extraction method (Fig. 7) is particularly crucial in brain tumor segmentation, in which both tumor homogeneity and precise boundary delineation are essential for clinical accuracy.

The integration of the TSHAO module and iterative attention mechanisms further enhances the segmentation performance. These components enable progressive feature refinement and improved spatial coherence and contextual understanding. Visualization of the decoder features (Fig. 8) validated this effect, demonstrating increasingly precise activation patterns around the tumor regions in successive layers of attention.

Furthermore, the optimal performance observed with a patch size of 16 × 16 and an encoder–decoder depth of 5–5 indicates a balance between fine detail preservation and overfitting mitigation (Table 6). This finding aligns with trends in medical image segmentation, emphasizing the significance of multiscale feature representation and hierarchical learning.

Clinical implications

The exceptional quantitative metrics achieved by DSIT-UNet (>95% mean IoU and mDice across all datasets) underscores its potential for clinical applications. Notably, the model maintained high recall values (96.55% for TCIA, 95.55% for BraTS 2019, and 96.44% for BraTS 2020), indicating robust tumor detection with minimal false negatives (Table 2). This is particularly valuable in clinical settings where missing tumor regions can significantly affect treatment planning and patient outcomes.

Saliency map analysis (Fig. 9) further enhances the clinical relevance of the model by providing interpretable decision-making insights. Strong activation along tumor boundaries and gradual intensity variations within tumor regions align with radiological expertise, reinforcing the trustworthiness of the model for AI-assisted medical diagnosis.

In addition, the computational efficiency of DSIT-UNet makes it a viable candidate for deployment in resource-constrained clinical environments (Table 3). With only 15.97M parameters, DSIT-UNet is significantly more efficient than transformer-based alternatives, such as TransUNet (177.91M) and DS-TransUNet (107.56M). Although the FLOP count is higher than that of traditional U-Net variants, substantial gains in segmentation performance justify this computational cost, given the critical nature of accurate tumor segmentation in clinical practice.

Comparative analysis with existing methods

DSIT-UNet consistently outperformed existing methods across diverse datasets, highlighting its robustness and generalizability. Conventional CNN-based models, such as U-Net and its variants (Res-UNet, Res-UNet++), struggle to capture complex tumor morphologies, as evidenced by their lower IoU and Dice scores. Although transformer-based models (TransUNet and DS-TransUNet) enhance global context modeling, they fall short of the boundary precision and computational efficiency.

By integrating the strengths of both paradigms, the DSIT-UNet mitigates these shortcomings through its dual-stream architecture and iterative attention mechanisms interms of rehabilitation. The qualitative results (Fig. 6) visually illustrate these improvements, showing more accurate tumor location and boundary delineation compared to SOTA methods. Feature visualization (Figs. 7 and 8) further elucidated the distinct contributions of each architectural component.

Limitations and future directions

Despite its promising performance, the DSIT-UNet has several limitations that warrant further exploration. Despite these promising results, this study had several limitations that warrant consideration. First, although DSIT-UNet achieved exceptional performance in brain tumor segmentation, its generalizability to other medical imaging tasks remains to be validated. Future studies should evaluate the performance of this model in diverse anatomical structures and imaging modalities to assess its broader applicability to medical image analysis.

Second, the increased computational complexity compared to conventional U-Net variants (385.05G FLOPs vs. 144.11G for U-Net) may present challenges for real-time applications or deployment on edge devices with limited computational resources. Further research could focus on model compression techniques, such as knowledge distillation or quantization, to reduce the computational overhead while maintaining segmentation accuracy.

Third, although our saliency map analysis provides some interpretability, more comprehensive explainability frameworks could enhance the transparency of the model for clinical adoption. Future studies could explore integrated attention visualization, rehabilitation strategies, feature attribution methods, or counterfactual explanations to provide more detailed insights into the model's decision-making process.

Finally, the current evaluation focuses primarily on segmentation accuracy, with a limited assessment of the robustness of the model to domain shifts or image-quality variations. Future studies should investigate the performance of the model under varying acquisition parameters, scanner types, and image quality conditions to ensure a reliable performance in real-world clinical settings.

Conclusion

DSIT-UNet represents a significant advancement in brain tumor segmentation, achieving state-of-the-art performance while maintaining a reasonable computational efficiency. The dual-stream architecture with mixed kernel configurations coupled with iterative attention mechanisms and the TSHAO module effectively addressed the challenges of accurate tumor boundary delineation and regional consistency. Comprehensive ablation studies have validated the contribution of each architectural component and have offered valuable insights for future medical image segmentation research.

The high recall values of the model, interpretable decision-making process visualized through saliency maps, and balanced trade-off between performance and computational complexity make DSIT-UNet a promising tool for assisting radiologists in the assessment of brain tumors and the planning of treatment and rehabilitation. These advancements could improve the diagnostic accuracy and workflow efficiency in neuro-oncology.

Future research should focus on extending the applicability of the model to diverse medical imaging tasks, optimizing the computational efficiency for deployment in resource-limited environments, enhancing explainability frameworks, and evaluating the robustness to domain shifts and image quality variations. These efforts will further bridge the gap between algorithmic innovation and clinical implementation, ultimately benefiting patient care through more precise and reliable tumor segmentation with conventional therapies and rehabilitation.

Data availability

All data generated or analyzed during this study are included in this published article.

Received: 11 December 2024; Accepted: 11 April 2025
Published online: 22 April 2025

References

1. Ghaffari, M., Sowmya, A. & Oliver, R. Automated brain tumor segmentation using multimodal brain scans: A survey based on models submitted to the brats 2012–2018 challenges. *IEEE Rev. Biomed. Eng.* **13**, 156–168 (2019).
2. Louis, D. N. et al. The 2016 world health organization classification of tumors of the central nervous system: A summary. *Acta Neuropathol.* **131**, 803–820 (2016).
3. American Cancer Society. American cancer society (2024). Accessed 2024-07-15.
4. Shukla, G. et al. Advanced magnetic resonance imaging in glioblastoma: A review. *Chin. Clin. Oncol.* **6**, 40–40 (2017).
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, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, 234–241 (Springer, 2015).
6. Avazov, K., Mirzakhalilov, S., Umirzakova, S., Abdusalomov, A. & Cho, Y. Dynamic focus on tumor boundaries: A lightweight u-net for mri brain tumor segmentation. *Bioengineering* **11**, 1302 (2024).
7. Vaswani, A. Attention is all you need. *Advances in Neural Information Processing Systems* (2017).
8. Liu, Z. et al. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022 (2021).
9. Jaegle, A. et al. Perceiver: General perception with iterative attention. In *International Conference on Machine Learning*, 4651–4664 (PMLR, 2021).
10. Chen, J. et al. Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021).
11. Cao, H. et al. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision*, 205–218 (Springer, 2022).
12. Cao, H. et al. A quantitative model based on clinically relevant mri features differentiates lower grade gliomas and glioblastoma. *Eur. Radiol.* **30**, 3073–3082 (2020).
13. Pereira, S., Pinto, A., Alves, V. & Silva, C. A. Brain tumor segmentation using convolutional neural networks in mri images. *IEEE Trans. Med. Imaging* **35**, 1240–1251 (2016).
14. Jlassi, A., ElBedoui, K. & Barhoumi, W. Brain tumor segmentation of lower-grade glioma across mri images using hybrid convolutional neural networks. In *ICAART (2)*, 454–465 (2023).
15. Liu, Z. et al. Canet: Context aware network for brain glioma segmentation. *IEEE Trans. Med. Imaging* **40**, 1763–1777 (2021).
16. Liu, X., Hou, S., Liu, S., Ding, W. & Zhang, Y. Attention-based multimodal glioma segmentation with multi-attention layers for small-intensity dissimilarity. *J. King Saud Univ. Comput. Inf. Sci.* **35**, 183–195 (2023).
17. Zhang, X., Hu, Y., Chen, W., Huang, G. & Nie, S. 3d brain glioma segmentation in mri through integrating multiple densely connected 2d convolutional neural networks. *J. Zhejiang Univ. Sci. B* **22**, 462–475 (2021).
18. Wu, S., Wang, G., Tang, P., Chen, F. & Shi, L. Convolution with even-sized kernels and symmetric padding. *Advances in Neural Information Processing Systems* **32** (2019).
19. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**, 1735–1780 (1997).
20. Dosovitskiy, A. et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020).
21. Wang, X., Girshick, R., Gupta, A. & He, K. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7794–7803 (2018).
22. Wu, H. et al. Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22–31 (2021).
23. Xu, W., Xu, Y., Chang, T. & Tu, Z. Co-scale conv-attentional image transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9981–9990 (2021).
24. Lee, Y., Kim, J., Willette, J. & Hwang, S. J. Mpvit: Multi-path vision transformer for dense prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7287–7296 (2022).
25. Guo, J. et al. Cmt: Convolutional neural networks meet vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12175–12185 (2022).
26. Chen, Q. et al. Mixformer: Mixing features across windows and dimensions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5249–5259 (2022).
27. Menze, B. H. et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE Trans. Med. Imaging* **34**, 1993–2024 (2014).
28. Bakas, S. et al. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Sci. Data* **4**, 1–13 (2017).
29. Mazurowski, M. A. et al. Radiogenomics of lower-grade glioma: Algorithmically-assessed tumor shape is associated with tumor genomic subtypes and patient outcomes in a multi-institutional study with the cancer genome atlas data. *J. Neurooncol.* **133**, 27–35 (2017).
30. Bakas, S. et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. arXiv preprint arXiv:1811.02629 (2018).
31. Buda, M., Saha, A. & Mazurowski, M. A. Association of genomic subtypes of lower-grade gliomas with shape features automatically extracted by a deep learning algorithm. *Comput. Biol. Med.* **109**, 218–225 (2019).
32. Fan, D.-P. et al. Pranet: Parallel reverse attention network for polyp segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 263–273 (Springer, 2020).
33. Jha, D., Riegler, M. A., Johansen, D., Halvorsen, P. & Johansen, H. D. Doubleu-net: A deep convolutional neural network for medical image segmentation. In *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*, 558–564 (IEEE, 2020).
34. Lin, A. et al. Ds-transunet: Dual swin transformer u-net for medical image segmentation. *IEEE Trans. Instrum. Meas.* **71**, 1–15 (2022).
35. Xiao, X., Lian, S., Luo, Z. & Li, S. Weighted res-unet for high-quality retina vessel segmentation. In *2018 9th International Conference on Information Technology in Medicine and Education (ITME)*, 327–331 (IEEE, 2018).
36. Jha, D. et al. Resunet++: An advanced architecture for medical image segmentation. In *2019 IEEE International Symposium on Multimedia (ISM)*, 225–2255 (IEEE, 2019).
37. Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F. & Adam, H. Encoder–decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 801–818 (2018).

38. Simonyan, K., Vedaldi, A. & Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034 (2013).

Acknowledgements
The authors thank the Bio-Imaging Research Lab, Department of Biomedical Engineering, Islamic University, Bangladesh for their support. Also, the authors extend their appreciation to the King Salman center for Disability Research for funding this work through Research Group no KSRG-2024-117 and the researchers supporting project (SRG-242277) of the Ministry of Science and Technology, Dhaka, Bangladesh.

Author contributions
Shakib Al Hasan: Conceptualization, Formal analysis, Methodology, Software, Data curation, Validation, Visualization, Writing—Original Draft, Review & Editing. S M Mahim: Formal analysis, Methodology, Software, Data curation, Validation, Visualization, Writing—Review & Editing. Md Emamul Hossen: Formal analysis, Methodology, Software, Visualization, Writing—Review & Editing. Md Olid Hasan: Data Curation, Validation, Visualization, Writing—Review & Editing. Md Khairul Islam: Formal analysis, Supervision, Investigation, Validation, Writing—Review & Editing. Patrizia Livreri: Writing—Review & Editing. Salah Uddin Khan: Formal analysis, Methodology, Validation, Software, Writing—Review & Editing. Mohammad Alibakhshikenari: Writing—Review & Editing. Md Sipon Miah: Project Administration, Writing—Review & Editing.

Declarations

Competing interests
The authors declare that they have no conflict of interest.

Additional information
Correspondence and requests for materials should be addressed to M.K.I., P.L., M.A. or M.S.M.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.