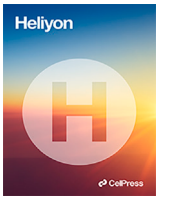
Heliyon 9 (2023) e22412

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

HUT: Hybrid UNet transformer for brain lesion and tumour segmentation

Wei Kwek Soh, Hing Yee Yuen, Jagath C. Rajapakse *

Nanyang Technological University Biomedical Informatics Lab, Block NS4-04-33 50 Nanyang Avenue, Singapore, 639798, Singapore, Singapore

A R T I C L E I N F O

Dataset link: https://www.med.upenn.edu/cbica/brats2020/data.html

Dataset link: https://fcon_1000.projects.nitrc.org/indi/retro/atlas.html

Keywords:
Brain tumour
Brain lesions
Multimodal MRI
Single-modal MRI
Self-supervised segmentation
Vision transformer

A B S T R A C T

A supervised deep learning network like the UNet has performed well in segmenting brain anomalies such as lesions and tumours. However, such methods were proposed to perform on single-modality or multi-modality images. We use the Hybrid UNet Transformer (HUT) to improve performance in single-modality lesion segmentation and multi-modality brain tumour segmentation. The HUT consists of two pipelines running in parallel, one of which is UNet-based and the other is Transformer-based. The Transformer-based pipeline relies on feature maps in the intermediate layers of the UNet decoder during training. The HUT network takes in the available modalities of 3D brain volumes and embeds the brain volumes into voxel patches. The transformers in the system improve global attention and long-range correlation between the voxel patches. In addition, we introduce a self-supervised training approach in the HUT framework to enhance the overall segmentation performance. We demonstrate that HUT performs better than the state-of-the-art network SPiN in the single-modality segmentation on Anatomical Tracings of Lesions After Stroke (ATLAS) dataset by 4.84% of Dice score and a significant 41% in the Hausdorff Distance score. HUT also performed well on brain scans in the Brain Tumour Segmentation (BraTS20) dataset and demonstrated an improvement over the state-of-the-art network nnUnet by 0.96% in the Dice score and 4.1% in the Hausdorff Distance score.

1. Introduction

Malignant brain tumours are life-threatening and the 10th leading cause of death worldwide. An estimated 251,329 people died from primary cancerous brain and Central Nervous System tumours in 2020 [1]. Magnetic Resonance Imaging (MRI) is widely used to detect and segment brain lesions and tumours because of its high contrast. Segmentation of brain lesions or tumours is one of the many essential tasks for early and late diagnostic and therapeutic intervention. A clot in a blood vessel that supplies the brain with oxygen can cause an ischemic stroke. It occurs when the brain lacks these important substances. This process can be life-threatening and can cause permanent brain damage or death. In 2020, an estimated 3.48 million deaths globally due to ischemic stroke [2]. Early diagnosis of ischemic stroke can help identify treatment options and improve patient outcomes. Medical experts still use manual segmentation of stroke lesions from MRI scans to identify areas and boundaries of damaged brain tissue in diagnosing stroke.

* Corresponding author.
E-mail address: asjagath@ntu.edu.sg (J.C. Rajapakse).

https://doi.org/10.1016/j.heliyon.2023.e22412
Received 14 August 2023; Received in revised form 23 September 2023; Accepted 10 November 2023

Available online 17 November 2023

Manual segmentation of brain abnormalities is a time-consuming process that requires expertise and can produce inconsistent results depending on the expert performing it. Therefore, automatic brain tumour and lesion segmentation using algorithms may offer a more efficient and objective approach to diagnosing and treating the medical condition. Deep learning methods have contributed to significant advancements in automated brain segmentation over time. In particular, supervised deep learning methods have yielded tremendous progress and performance over existing machine learning techniques. UNet [3] is one of the pioneering deep-learning methods in biomedical image segmentation. The architecture of the UNet comprises downsampling and upsampling convolutional layers with in-between skip-connection to improve the gradient backpropagation during the network training. The model can capture multi-resolution features by expanding and contracting spatial resolution and depth structures, which help in image segmentation performance. The state-of-the-art method like nnUnet [4] introduces a set of training enhancements to the existing 3D UNet architecture. It encompasses deeper loss functions, backpropagation at the encoder branches, and additional data augmentation. During the regular training of brain tumour segmentation, the labels provided for the model are "edema," "non-enhancing tumour and necrosis," and "enhancing tumour." In nnUNet, region-based training was used. The segmentation is evaluated based on three partially overlapping regions: the whole tumour (which includes all three classes), the tumour core (comprised of non-enhancing and necrosis and enhancing tumour), and the enhancing tumour. Previous research has shown that focusing on optimizing these regions rather than the individual classes can lead to improved performance on the segmentation task. The nnUnet model achieved better results than the previous state-of-the-art approaches by combining the predictions of multiple trained models on different folds of the dataset.

As for the task of lesion segmentation, [5] demonstrated an approach to achieve state-of-the-art segmentation for the ATLAS R1.2 dataset. Their system uses two networks to make predictions. One network maps the input to a high-dimensional embedding space at twice the input's resolution and the other produces a confidence map using "subpixel" predictions. Four predictions within a 2×2 neighbourhood represent an output pixel segmentation. The output class for each pixel segmentation is obtained by using a learnable downsampler to predict the contribution of each subpixel prediction in a local region corresponding to the pixel in the original resolution, which avoids the need to use hand-crafted downsampling techniques such as bilinear or nearest neighbour interpolation.

Besides highly successful convolutional networks, Transformer architecture has recently shown notable performance in image classification and segmentation. In the paper [6], the authors introduced Transformers as self-attention networks for better Natural Language Processing (NLP) tasks. As a result, Bi-directional Encoder Representations from Transformers (BERT) [7] became a popular pre-trained language model using an unlabelled text corpus. It performs well in various NLP tasks compared to models trained for individual tasks. Only recently, the Transformer has emerged as a learning candidate for computer vision. The architecture of the Transformer is naturally suited for sequences. However, the problem with vision is that images do not appear in a sequence. The workaround to using a Transformer in vision is to create patches of the images and map the patches to learnable or fixed latent vectors, arranged as an input sequence to the Transformer.

Vision Transformer (ViT) has been adopted in the applications of medical imaging analysis in recent studies. It works well for tumour-type classification of ultrasound images [8]. In brain tumour segmentation, there has been an increase in research using ViT [9], [10], [11], [12], [13], [14]. Nevertheless, CNN-based UNet [3] remains a strong contender for medical imaging segmentation tasks. Therefore, researchers are now looking at hybrid architectures combining UNet and Transformers. In [11], the authors used the Swin Transformer from [15] in the form of a UNet. They addressed the issue of computing resources due to the quadratic complexity of Transformer architecture when the input dimension becomes large. It utilizes a hierarchical structure to reduce the complexity by merging the patches in subsequent layers. In addition, introducing a shifting window mechanism increases the receptive field. However, the downside of Swin-Unet is that it only operates with 2D scans since it is pre-trained on the ImageNet dataset. Pure Transformer networks are generally not data-efficient, requiring more data resources to train [16]. Tang et al. [17] introduced a 3D Swin architecture, Swin-Unetr, with self-supervised learning to improve the performance in brain tumour segmentation. Authors of [18], [19], [12] proposed hybrid architectures referred to as Mixed-Transformer UNet (MT-UNet), Transformer Brain Tumour Segmentation (TransBTS) and Unet Transformer (Unetr), respectively. They took small feature representations from CNN. They implemented a stack of transformers for the bottleneck to reduce the footprint of the Transformer and reduce the overall complexity.

Multi-modal images provide brain information variability and image diversity that aids in improving diagnosis and segmentation performances. The publicly available Brain Tumour Image Segmentation Benchmark 2020 (BraTS 2020) dataset [20], [21], [22] contains 3D MRI scans and includes four modalities for each patient data. The four modalities are (a) native T1-weighted(T1), (b) T2-weighted(T2), (c) post-contrast T1-weighted(T1ce), (d) T2 Fluid Attenuated Inversion Recovery (Flair). These modalities provide complementary information in different imagery and intricate details essential for accurate tumour detection and segmentation. The four modalities in the Brain Tumour Segmentation (BraTS) 2020 dataset are produced through different acquiring methods. The FLAIR modality is very sensitive to tumours and differentiates between Cerebrospinal Fluid (CSF) and tumour. T1ce images are produced by improving the contrast of the T1 images using a contrast enhancement agent. With these imagery differences and various distinctive features and information displayed in each modality, T2 and FLAIR modality images are most suitable for detecting peritumoral edema tumours. In contrast, T1 and T1ce modality images are best fitted to classify the tumour core. Therefore, it is apparent that each modality captures distinct brain anatomy and tumour characteristics. Complete modality data provides pervasive and rich data crucial for the segmentation performance of models used for these tasks.

In contrast, we only utilize a single-modality of MRI scans to segment brain lesions, typically the T1-weighted MRI scan from the Anatomical Tracings of Lesion After Stroke (ATLAS) R1.2 dataset. Damaged or dead brain tissue caused by an ischemic stroke is very different from healthy brain tissue. This can affect the strength of the signal produced by the tissue and cause it to appear darker on the MRI image. One challenge in using convolutional neural network (CNN) architectures for tumour and lesion segmentation is the difficulty in accurately segmenting small lesions. CNNs use techniques such as max pooling and strided convolution to acquire global features of the input image. Still, reducing spatial resolution can result in the loss of detailed information about small local

structures. Hence, current CNN-based methods do not perform well on small lesions and can lead to unclear segmentation between healthy brain tissue and lesions or even missed segmentation.

According to [16], the ViT outperforms its CNN counterpart in terms of accuracy and computational efficiency of image classification. ViT, like the original Transformer, aims to capture long-range and short-range correlations in sequence data using a self-attention mechanism. The images are processed into patches and arranged to the ViT model as sequence data. We establish a two-fold approach to exploit the inter-correlation between the modalities and the intra-correlation between the voxels. First, we incorporate the ViT with convolution layers to attend to the lesion and tumour anomalies. Second, we present a self-supervised methodology to learn latent features and improve the overall dice score.

Our approach differs from existing hybrid systems such as Unetr, TransBTS [19], TransUnet [13], and STHarDNet [14]. Unetr utilizes CNN layers at the output of the skip connections and concatenates the output sequence representation from the upsampled CNN layers with a decoder similar to UNet. TransBTS places the transformer at the bottleneck of the UNet architecture. At the same time, STHarDNet adds a Swin Transformer at the first skip connection of UNet and concatenates its output at the second layer before the final layer of the UNet decoder. TransUnet is similar to TransBTS but has an additional downsampling CNN layer at the transformer's output. In contrast, our architecture takes two sizes of patches at the input and multiplies the attention map from the output of the cross-transformer at the UNet decoder. Additionally, we utilize self-supervised training for the CLS token at the output of the cross-transformer to enhance performance.

In summary, we make the following key contributions in this paper:

1. The hybrid UNet Transformer (HUT) is an advanced solution for brain tumour segmentation as demonstrated by its performance on the BraTS20 dataset consisting of four imaging modalities (T1w, T2w, Flair, T1ce). The HUT system combines the Vision Transformer pipeline and the data-efficient UNet pipeline, resulting in faster training convergence and improved overall performance.
2. Adding upsampling and layering to the HUT achieves better performance on the single-modality (T1w) ATLASR12 dataset and improves its ability to capture smaller lesions.
3. HUT enables self-supervision of transformer networks on top of the supervised training of the UNet network and attention maps of the Vision Transformer.

2. Methodology

2.1. Hybrid UNet transformer (HUT) architecture

The UNet architecture is the best choice for medical imaging segmentation due to its exceptional performance. It uses convolutional layers that exploit the local correlation between pixels via the kernels, provide an inductive bias to the system, and increase the convergence rate of learning. The Transformer offers a long-range relationship between the tokens represented mainly by image patches. However, transformers are not data-efficient and require large datasets for the training to converge appropriately. We also know that annotated data is costly and scarce in medical imaging. In light of this limitation, we introduce a hybrid network HUT that incorporates the merits of both convolution layers and Transformers [18], [19]. However, our approach is distinctly different from previous work. We implemented the Transformer at the last few skip connections rather than at the bottleneck layer, providing a better receptive field. Moreover, we extracted the probability vectors from the final classification layers of the cross-transformer for self-supervising learning.

Fig. 1 illustrates the overall architecture of HUT for brain tumour and lesion segmentation. This architecture has two pipelines: the UNet pipeline (UNP) and the Vision Transformer pipeline (VTP). First, we infuse the Transformers blocks into the UNet structure via skip connections. Then, within the Transformers module, we instantiate a voxel embedding, a local Transformer, a position embedding, a global Transformer and finally, a voxel decoder. The local Transformer acts on the local voxel patches, whereas the global Transformer acts on the overall image. Adopting a hybrid architecture can address the voxels' local and global correlation issues. It also achieves much faster convergence than training a pure Transformer-based UNet.

The traditional UNet has better convergence, mainly due to inductive bias from the CNN architecture. The ViT picks up the salient relationship between the voxels. In our proposed architecture, the Transformer pipeline operates parallel to the UNet pipeline and multiplies the sub-outputs of the cross-transformer to the decoder layers of the UNet. Intuitively, they help provide appropriate attention to critical regions at different resolutions.

The HUT architecture is introduced to address the shortcomings of convolutional networks and transformer networks. The convolutional network of the UNet is data-efficient and generalises image features well due to its inductive bias. The transformer network uses the self-attention mechanism to correlate long-range dependencies between image tokens. However, the transformer architecture uses large amounts of data to generalise well. Therefore, the HUT architecture incorporates the VTP parallel to the UNP to overcome these limitations. By doing so, we observe that VTP is trained more efficiently while UTP can capture long-range relationships between different patches.

2.2. The UNet pipeline (UNP)

The nnUnet training framework inspires the UNet pipeline with deeper supervision of region-based labels [4]. The UNet pipeline comprises the conventional UNet structure. The difference between the two is in the training, i.e., the supervision of the labels. The

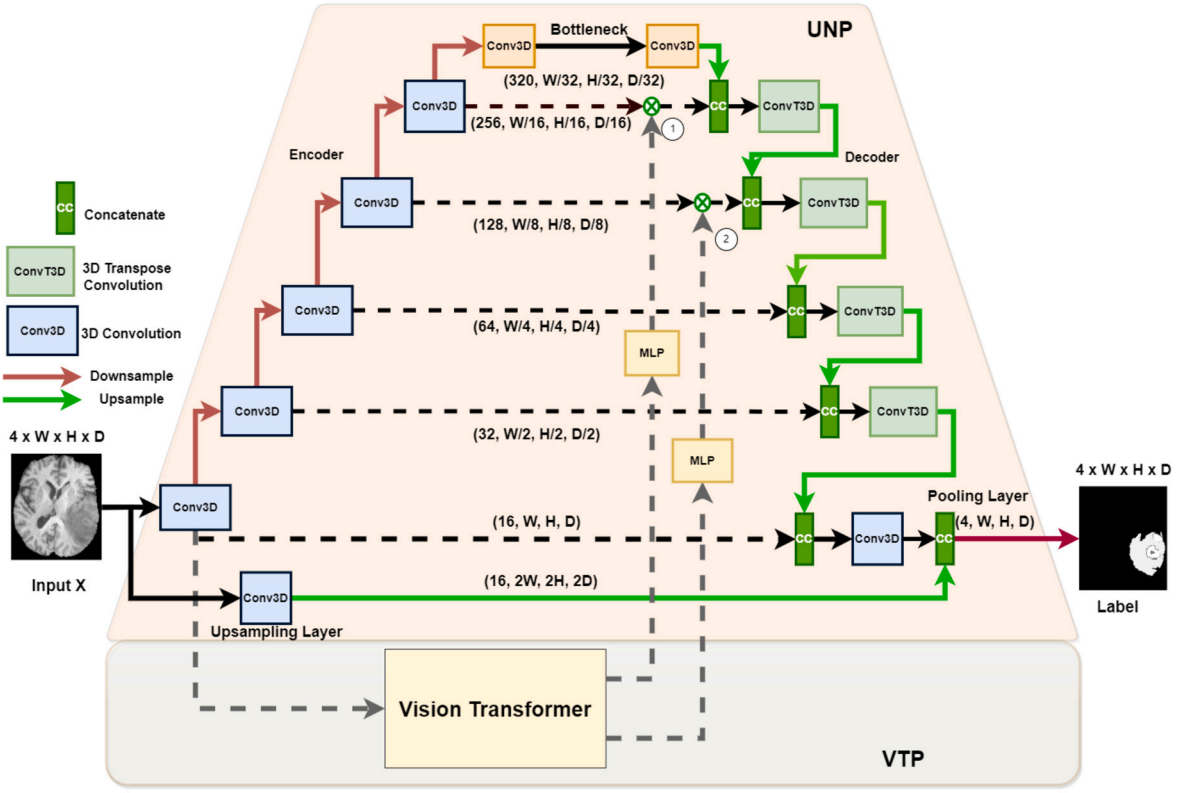


Fig. 1. The HUT architecture consists of two separate pipelines: the UNet pipeline (UNP) and Vision Transform Pipeline (VTP).

decoder is trained with multiple resolutions of the labels, as explained in the next section on the training of HUT. In addition to this architecture, we proposed introducing extra upsampling and pooling layers to the first skip connection, as illustrated in Fig. 1. It increases the receptive field, improves the ability to capture minute details, and more effectively captures small anomalies such as lesions. However, this approach typically does not provide additional gain in brain tumour segmentation compared to brain lesion segmentation since the tumours are much larger than the lesions.

The UNP leverages the merit of convolution layers on its inductive biases and allows the other transformer pipeline to converge much faster. As a result, the training of the VTP becomes more data-efficient compared to the traditional training of the transformer, which requires a large amount of data.

2.3. The vision transformer pipeline (VTP)

The Transformer architecture has been very effective and has performed best on many natural language processing tasks. In contrast, a Vision Transformer (ViT) is a variant of the Transformer designed explicitly for identifying objects and patterns in images. It converts an image into a series of patch tokens by dividing the images into smaller patches and projecting each patch into a set of tokens through a linear transformation. An extra token of the Transformer architecture used for NLP tasks is the CLS (Classification token). We typically add and summarise it at the start of the input sequence, enabling the Transformer to perform classification tasks like sentiment analysis or text classification [7]. Utilizing the hidden states of every token in the input sequence, the Transformer computes the representation of the CLS token. It uses the representation as a single feature vector to provide a prediction through a classifier layer. The token makes it possible for the Transformer to gather data from the entire input sequence in a condensed form and base predictions on that data. Instead of only considering a portion of the sequence or individual tokens, using CLS token is constructive when the entire input sequence is required to make a prediction. Some variant of the ViT uses the CLS token for the final classification task.

The VTP comprises two sub-layers: a self-attention layer and a feedforward layer. The self-attention layer enables the network to determine the significance of different elements in the input sequence and incorporate this information into its input representation. The ViT includes position embedding in each token to include the vital position information for many vision tasks. This is necessary because the self-attention mechanism in the Transformer encoder does not consider the position of the input elements. The VTP aggregates the information from other tokens into the CLS token at the penultimate stage before a classification feedforward layer finally maps it to the output.

For VTP, as illustrated in Fig. 2, we adopt two parallel transformer structures to handle small and large image patches of two different resolutions [23]. Each voxel in the images is mapped to an embedding: small patch attention (*SPA*) operates on sub-regions

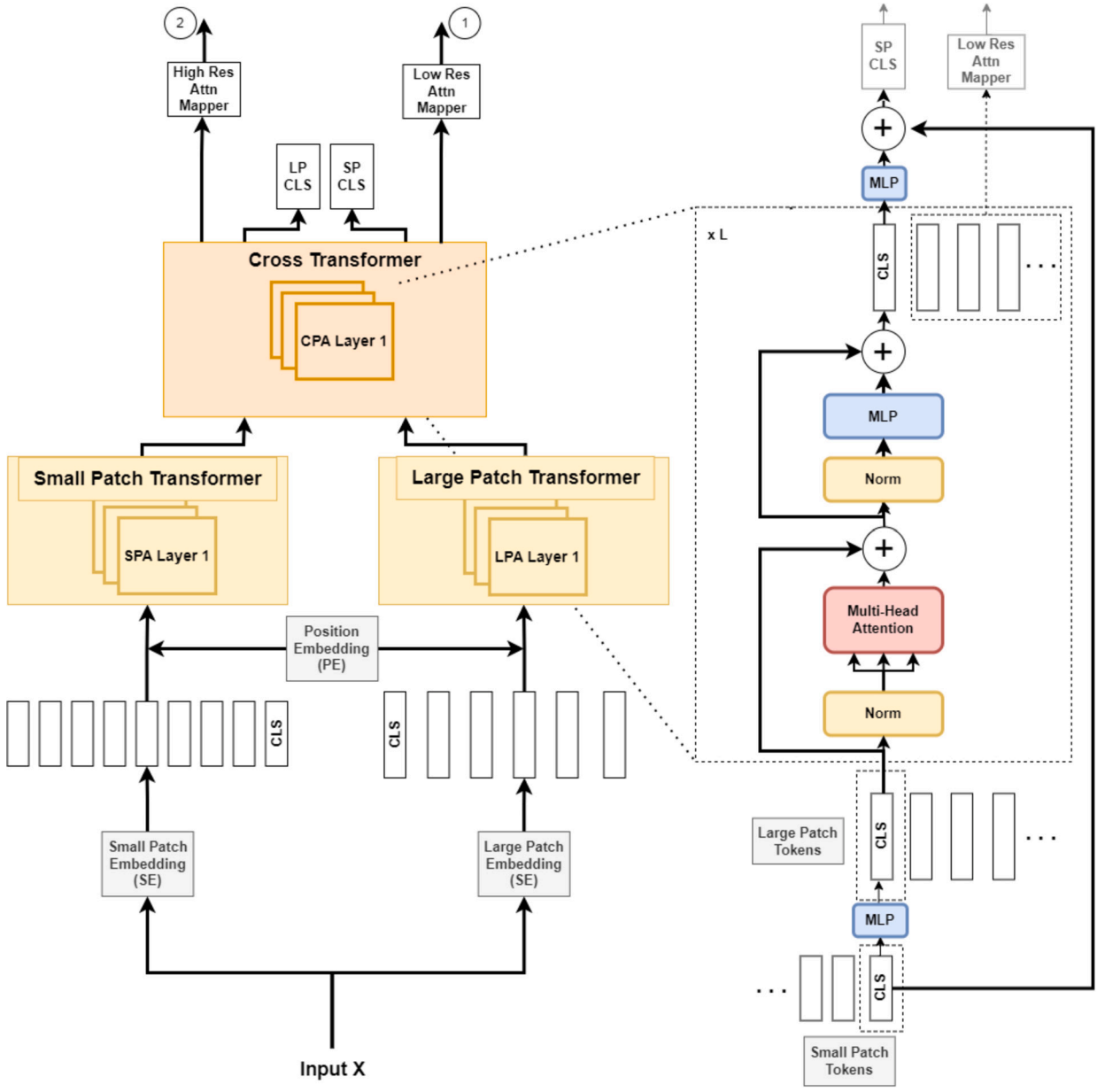


Fig. 2. The Vision Transformer pipeline (VTP) consists of dual-branch transformers for the small patches and large patches. The outputs of both transformers are aligned as a sequence at the last stage of cross-transformer to correlate the tokens.

of voxel embedding to ensure self-attention correlation within its small patch group, and the large patch attention (*LPA*) works on the coarser resolution. f_s and f_l are the linear network mapper functions that map to the same dimension and R_s as the residual operation such that $R_s(f(X)) = f(X) + X$. We denote X as the input voxel data such that $X \in \mathbb{R}^{W \times H \times D \times C}$, where W, H, D are the dimensions of input, and C is the channel length.

Embedded small patch output $Y^{SA} \in \mathbb{R}^{(W/p)\times(H/p)\times(D/p)\times C_S}$, with dimensions $W/p, H/p, D/p$, are the embedded voxel patches from embedding function with patch size p and C_S is the new embedding channel length. Similarly, embedded large patch output $Y^{LA} \in \mathbb{R}^{(W/k)\times(H/k)\times(D/k)\times C_L}$, with dimensions $W/k, H/k, D/k$, are the embedded voxel patches from embedding function with patch size k with $k > p$, and C_L is the new embedding channel length.

We write the operation of the HUT architecture as follows:

$$\begin{aligned} S &= \mathrm{SE}(X) + \mathrm{PE} \\ Y^{SA} &= f_s(R_s(\mathrm{LN}(\mathrm{SPA}(R_s(\mathrm{LN}(S)))))) \end{aligned} \tag{1}$$

$$\begin{aligned} L &= \mathrm{LE}(X) + \mathrm{PE} \\ Y^{LA} &= f_l(R_s(\mathrm{LN}(\mathrm{LPA}(R_s(\mathrm{LN}(L)))))) \end{aligned} \tag{2}$$

where S and L are the output of voxel patch embedding of small and large patches, respectively. Y^{SA} and Y^{LA} denote the outputs from SPA and LPA blocks, respectively. We denote the learnable small and large voxel patch embedding and position encoding function as SE, LE and PE respectively, and LN as layer normalisation operation.

$$\begin{aligned} W &= \texttt{Concatenate}(Y^{SA}, Y^{LA}) \\ Z &= R_s(f(\mathrm{LN}(\mathrm{CPA}(R_s(\mathrm{LN}(W)))))) \end{aligned} \tag{3}$$

where W results from the concatenated output from the two transformers. Z denotes the output from the cross-transformer block, which comprises the cross-patches attention (CPA) function.

The cross-attention is an efficient and fast way of combining the two transformers [24], which involves merging the CLS token from one branch with the patch tokens from another. To make the most of this fusion and efficiently combine multi-resolution features, we use the CLS token at each pipeline as a conduit for exchanging information between the patch tokens of the other branch and then bring that information back to its pipeline.

The CLS token, trained to gain abstract information across all the patch tokens in its branch, can benefit from correlating with the patch tokens in another branch. For instance, after combining with the patch tokens from the other branch, the CLS token can share the newly learned information with its patch tokens in the next transformer encoder. Adding additional context and insights helps improve each patch token's representation.

To obtain the CLS token from the larger patch transformer, we concatenate it with the smaller patch tokens after projecting it to the same dimension. This combined set is then passed through the attention pipeline. This process also allows the CLS token from the smaller patch transformer to receive information from the larger patch tokens. Our work uses the softmax function to map the two final CLS tokens with projection headers to a fixed set of classes. We also add self-supervised training by comparing the probability distributions produced by the two CLS tokens.

2.4. The training of HUT

Training of HUT comprises supervised learning of the encoder-decoder of UNP and the self-supervised learning of the VTP, regularising the latent vectors of the CLS tokens by correlating the output probabilities. We enforced a loss function at the bottleneck during the training to encourage the encoder network to learn significant latent variables. The loss function is a cross-entropy loss function. The overall cost function consists of multiple components, which are described below.

2.4.1. Supervised loss functions

As shown in Fig. 3, HUT is trained with deeper supervision at scaled versions of the ground truth scan [4]. The gradient from the loss is backpropagated simultaneously at each branch during the training iteration.

In the supervised setting, the segmentation loss at the decoder comprises a weighted sum of the cross-entropy loss and soft dice loss. The cross-entropy loss matches the probability distribution p_c^i of the ground truth of pixel i and class label c and the probability of predicted output q_c^i. The cross-entropy loss $\mathcal{L}_{CE}$ is written as:

$$\mathcal{L}_{CE} = -\frac{1}{N}\sum_c \sum_i p_c^i log(q_c^i) \tag{4}$$

The soft dice loss, $\mathcal{L}_{Dice}$ uses the dice score of probabilities at the output of the softmax function of the network and is given by:

$$\mathcal{L}_{Dice} = \frac{1}{N}\sum_c 1 - \frac{2\sum_i q_c^i p_c^i}{\sum_i (q_c^i + p_c^i)} \tag{5}$$

where q_c^i is the probability of predicted class, p_c^i is the probability of actual class at pixel i and N denotes the number of batches.

2.4.2. Self-supervised loss function

We impose self-supervised training for HUT learning by matching the two output probability distributions from the CLS tokens. The self-supervised loss function is chosen as the cross-entropy loss to match the probability distribution p_{CLS} of the small patch CLS token and the probability q_{CLS} of the large patch CLS token. The cross-entropy loss $\mathcal{L}_{SS}$ between the probability output p_{CLS} from a small patch CLS token and the probability output q_{CLS} from a large patch CLS token is written as:

$$\mathcal{L}_{SS} = -\frac{1}{N}\sum_j p_{CLS}^j log(q_{CLS}^j) \tag{6}$$

where j is the j-th element of the softmax output of the projection layer after the CLS token.

2.4.3. Overall loss function

The training of HUT is end-to-end with the overall loss, $\mathcal{L}$ for the segmentation network represented as:

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda_{Dice}\mathcal{L}_{Dice} + \lambda_{SS}\mathcal{L}_{SS} \tag{7}$$

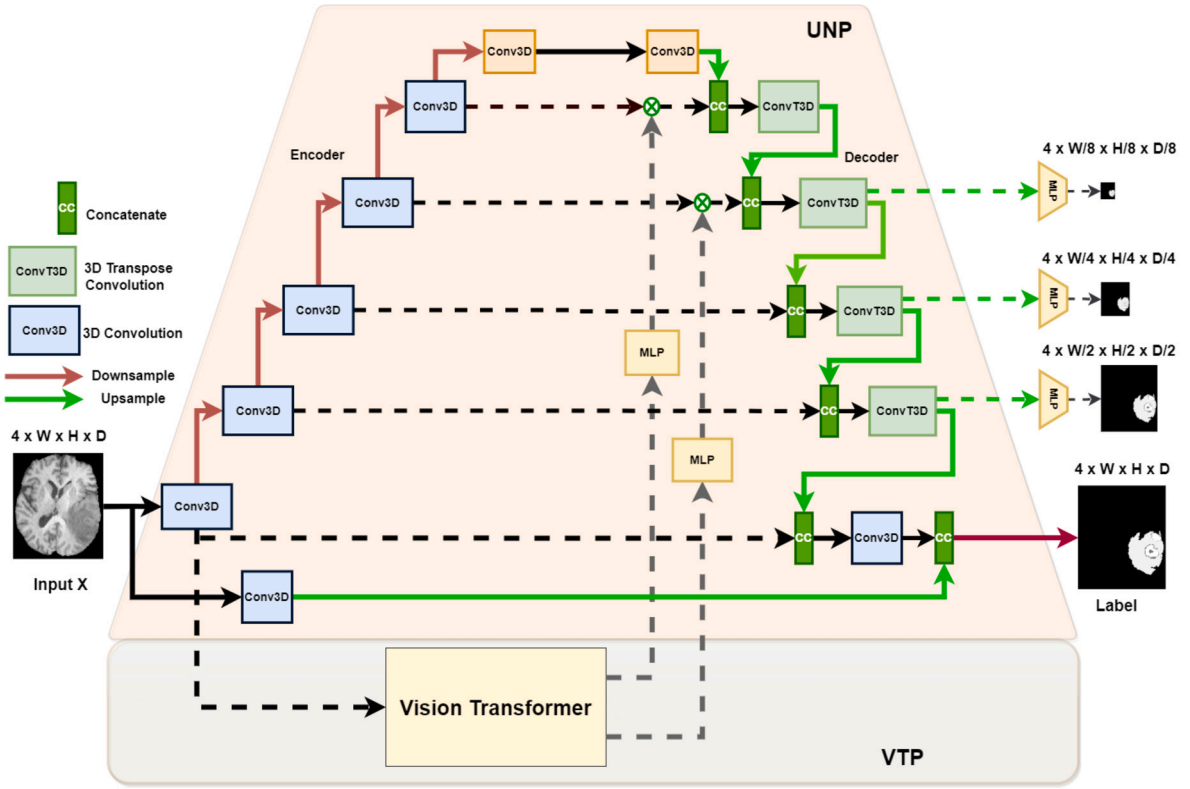


Fig. 3. The training of HUT consists of multiple scaled ground truth labels for Brain Tumour Segmentation.

where λ_{Dice} and λ_{SS} are the weighting factors for the dice and self-supervised losses, and from the experiments, the λ_{Dice} is empirically found as 0.5 for the BraTS segmentation task and 0 for the ATLAS lesion segmentation task. For λ_{SS}, it works empirically better with the value of $1e-5$.

3. Experiments and results

In this section, we demonstrate the performance of HUT on two datasets, namely, BraTS2020 and ATLASR1.2. We compared the proposed system's performance with state-of-the-art methods. The BraTS2020 dataset has four modalities imaging brain tumours, and the ATLASR12 dataset has a single modality imaging ischemic stroke lesions. The experiments were carried out on a workstation with one Nvidia RTX3090 GPU. The batch size is a single brain image volume per iteration.

3.1. Datasets

3.1.1. Brain tumour segmentation dataset (BraTS 2020)

The BraTS 2020 dataset comprises 369 high-grade multimodal MRI 3D images of brain tumours (Neuroblastoma). The images are skull-stripped and registered. Each volume's width, height, and depth are 240 pixels, 240 pixels, and 155 pixels, respectively. Each scan contains four modalities, namely T1, T2, FLAIR and T1ce. A segmentation map of each volume consisting of three classes, peritumoral edema, necrotic and non-enhancing tumour core, and gadolinium-based enhancing tumour, were manually annotated and curated by expert radiologists. In the experiments, we evaluate the performance according to the class groups: the whole tumour, the tumour core, and the enhancing tumour. The whole tumour is a visible region that includes the non-enhancing tumour, enhancing tumour and edema. The tumour core is a malignant tumour region consisting of the non-enhancing and enhancing tumour. The intensity of the T1ce scans is higher for the enhancing tumour due to the introduction of the gadolinium-based agent. The size of the input volume is $192 \times 192 \times 160$ so as to reduce GPU memory while at the same time, capturing the full tumour.

3.1.2. Ischemic stroke lesion segmentation dataset (ATLAS R1.2)

We used the ATLAS R1.2 dataset for the ischemic stroke lesion segmentation task. The ATLAS dataset [25], [26] consists of 304 T1-weighted MRI scans of stroke patients with corresponding lesion annotations. The data were collected from 11 research locations worldwide and manually annotated to identify the stroke lesion. The scans were then processed for privacy by smoothing and defacing, leaving 239 patient scans. We cropped each 3D scan to a resolution of $160 \times 160 \times 192$ to focus on relevant portions of the image and reduce the GPU memory needed. To compare with the results in [5], we used the same random data split of the ATLAS

dataset the authors had evaluated, with 212 train and 27 test subjects. The ischemic stroke dataset contains very small lesions, which can make segmentation tasks difficult.

3.2. Performance measures

We use the Dice score and HD95 score as the evaluation metrics to evaluate the testing set. The Dice metric calculates the overlap between predicted segmentation and ground truth regarding volume. It is determined by comparing the sets of voxels representing the object of interest in the ground truth (A) and the predicted segmentation (B). The Dice Score is flexible to variations in the number of elements in each set, making it a useful metric for contrasting sets of various sizes. The metric defines the similarity coefficient as taking the intersection of A and B and dividing it by the union of A and B.

$$DSC = \frac{2|A \cap B|}{|A| + |B|} \tag{8}$$

The Hausdorff distance (HD) is a measure that compares the similarity between two sets of points by calculating the maximum distance, D, between them. It is the maximum distance from one set to the nearest point in the other set.

$$HD = \max\{\max_{a \in A} \min_{b \in B} D(a, b), \max_{b \in B} \min_{a \in A} D(b, a)\} \tag{9}$$

The 95% HD is a variant of the Hausdorff distance based on the 95th percentile of all the distances between the boundary points of the two sets, A and B. HD95 is robust to outliers because it is less sensitive to them than the Dice Score, which makes it a useful metric when the sets being compared include outliers or extreme values.

In addition, we employed three other metrics for the comparison between the methods. The three other metrics are Intersection over Union (IoU), Precision and Recall. The IoU is defined as the fraction of the area of overlap over the area of the union between the predicted and ground truth segmentation. IoU assigns equal importance to both true and false positives and negatives, but it emphasises the completeness of the output segmentation. In contrast, the Dice score emphasises the accuracy of the segmentation more. Precision measures the exactness of the positives or the ratio of true positive outcomes against all true positive and false positive cases. Recall is the ratio of true positive outcomes against all true positive and false negative cases. A low precision and high recall will imply that the model is too optimistic in segmenting the regions while producing too many false positives, leading to lower accuracy. Suppose a model has high precision and low recall. In that case, it suggests that the model is conservative in segmenting the regions, resulting in too many false negatives and under-segmentation.

3.3. Comparison of performances with existing methods on brain tumour segmentation

The z-axis of the brain image volume was trimmed to a depth of 160 pixels, which provided complete coverage of the tumour region of interest. Similarly, we reduced the width and height of the axial to 192 on each side to reduce the need for GPU memory. We added data augmentation to increase the diversity of the data by rotation, scaling and masking. We compared similar training routines such as Data augmentation of random affine of (0.6,1.5), and rotation of 25 degrees with a probability of 30% for the 295 subjects.

We split the dataset into training and testing sets and randomly sampled 80% of the total 369 subjects and 20% for testing. All networks were trained and tested with one brain volume batch size with the same dataset. The hyper-parameters of the models were empirically found through 5-fold cross-validation. The testing data of 74 subjects were consistently similar throughout all experiments. The mean and standard deviation of the metrics were computed from the test samples.

We compared with various segmentation methods like nnUnet [4], Generative Adversarial Network (GAN)-based Vox2Vox [27] and Transformer-based SwinUNet [11] network. The Vox2Vox architecture uses weighted soft dice and adversarial loss to train the generative network that produces the segmentation map. The discriminator of Vox2Vox differentiates whether the output is a real or fake segmentation map. The SwinUNet architecture uses pure Transformer blocks as the UNet backbone. It takes the input, maps by patch embedding and merging, and hierarchically downsamples and upsamples with a shifted windowing approach. This approach creates wider multiple-window attention and a better global receptive field. Table 1 compares Dice scores and Hausdorff distance between the SOTA methods at different amounts of annotated data.

We did not implement the ensembling method of nnUnet to have a fair comparison among the state-of-the-art (SOTA) methods. We also note that the ensembling generates multiple models based on a cross-validated dataset. In this way, we feel that the testing set may have leaked into the training and, as a result, created an optimistic outcome. All the segmentation networks used the exact weighting of soft dice loss and cross-entropy loss for the training.

Table 1, 2, 3, 4, 5 illustrates the comparison between mean and standard deviation of Dice score, HD95 score, IoU, Precision and Recall of the BraTS20 segmentation with various methods. Our HUT method improves the mean Dice score performance over the SOTA method, nnUNet, by 0.96%, as shown in Table 1. It also enhances the performance of the mean 95th percentile Hausdorff Distance score (HD95) by 4.1%, as shown in Table 2. It is slightly better in the segmentation of the tumour core and enhancing tumour but marginally worse in the whole tumour than nnUnet. Although Vox2Vox is a GAN-based method, it trains its network with an additional supervised reconstruction component which uses the ground truth label.

HUT method improves the mean Dice score performance of the brain tumour segmentation over Vox2Vox by 1.7% and the HD95 score by 22.3%. Pure Transformer architecture like SwinUnet has achieved good segmentation performance by leveraging the pre-training technique using the large ImageNet dataset. The primary motivation of this method is to train the pure Transformer since

Table 1
Comparison between mean and standard deviation (in parentheses) of Dice score of the tumour (whole, enhanced, and core) segmentation against the other methods like Vox2Vox, nnUNet and SwinUNet. SwinUNet[1] is pre-trained with a supervised ImageNet of 1000 classes.

Methods	Whole	Core	Enhanced	Overall
Vox2Vox [27]	0.893 (0.071)	0.822 (0.122)	0.761 (0.249)	0.825 (0.165)
Unetr [12]	0.855 (0.128)	0.758 (0.152)	0.728 (0.245)	0.781 (0.182)
SwinUnet[1] [11]	0.872 (0.102)	0.799 (0.136)	0.785 (0.202)	0.819 (0.153)
SwinUnet [11]	0.770 (0.176)	0.661 (0.184)	0.651 (0.222)	0.694 (0.195)
nnUnet [4]	**0.904** (0.072)	0.830 (0.127)	0.760 (0.254)	0.831 (0.169)
TransBTS [19]	0.886 (0.112)	0.818 (0.140)	0.730 (0.250)	0.811 (0.178)
HUT (Ours)	0.900 (0.086)	**0.836** (0.126)	**0.783** (0.239)	**0.840** (0.163)

[1] SwinUnet [11] is pre-trained with ImageNet dataset of 1000 classes.

Table 2
Comparison between mean and standard deviation (in parentheses) of HD95 score (in mm) of the tumour (whole, core, and enhanced) segmentation by HUT against the SOTA methods.

Methods	Whole	Core	Enhanced	Overall
Vox2Vox [27]	7.833 (13.940)	7.172 (12.810)	4.077 (6.639)	6.361 (11.583)
Unetr [12]	10.563 (14.958)	9.599 (13.955)	6.045 (12.904)	8.736 (13.964)
SwinUnet[1] [11]	10.314 (15.773)	9.004 (14.130)	9.844 (19.117)	9.721 (16.471)
SwinUnet [11]	17.074 (17.514)	14.415 (16.288)	9.702 (16.310)	13.730 (16.714)
nnUnet [4]	**5.576** (10.03)	**5.273** (7.017)	4.300 (7.111)	5.222 (8.172)
TransBTS [19]	6.789 (11.536)	6.679 (10.991)	5.864 (8.824)	6.444 (10.515)
HUT (Ours)	6.054 (10.494)	5.523 (8.206)	**3.235** (5.476)	**4.937** (8.316)

Table 3
Comparison between mean and standard deviation (in parentheses) of IoU of the tumour segmentation by HUT against the SOTA methods.

Methods	Whole	Core	Enhanced	Overall
Vox2Vox [27]	0.814 (0.105)	0.714 (0.157)	0.662 (0.243)	0.730 (0.178)
Unetr [12]	0.764 (0.157)	0.631 (0.172)	0.617 (0.236)	0.671 (0.192)
SwinUnet[1] [11]	0.786 (0.135)	0.684 (0.169)	0.627 (0.249)	0.699 (0.190)
SwinUnet [11]	0.652 (0.187)	0.519 (0.185)	0.464 (0.214)	0.545 (0.196)
nnUnet [4]	**0.831** (0.102)	0.726 (0.160)	0.662 (0.250)	0.740 (0.181)
TransBTS [19]	0.806 (0.130)	0.711 (0.166)	0.621 (0.242)	0.714 (0.185)
HUT (Ours)	0.827 (0.112)	**0.733** (0.159)	**0.688** (0.235)	**0.749** (0.176)

it is not as data-efficient as the CNN counterpart, which has an inductive bias to focus on image objects. HUT method improves the mean Dice score performance over pre-trained SwinUnet, and SwinUnet trained from scratch by 2.4% and 20.9%, respectively. As for the HD95 score, it improves by 48.5% and 63.5%, respectively.

Table 4
Comparison between mean and standard deviation (in parentheses) of Precision of the tumour segmentation by HUT against the SOTA methods.

Methods	Whole	Core	Enhanced	Overall
Vox2Vox	0.900	0.849	0.792	0.847
[27]	(0.095)	(0.125)	(0.247)	(0.169)
Unetr	0.874	0.769	0.738	0.793
[12]	(0.162)	(0.178)	(0.264)	(0.206)
SwinUnet[1]	0.851	0.797	0.753	0.800
[11]	(0.152)	(0.161)	(0.277)	(0.205)
SwinUnet	0.745	0.643	0.613	0.667
[11]	(0.226)	(0.218)	(0.283)	(0.244)
nnUnet	0.906	0.848	0.802	0.852
[4]	(0.106)	(0.146)	(0.253)	(0.179)
TransBTS	**0.930**	**0.902**	**0.868**	**0.900**
[19]	(0.117)	(0.135)	(0.243)	(0.174)
HUT (Ours)	0.895	0.853	0.802	0.850
	(0.111)	(0.133)	(0.256)	(0.178)

Table 5
Comparison between mean and standard deviation (in parentheses) of Recall of the tumour (whole, enhanced, and core) segmentation by HUT against the SOTA methods.

Methods	Whole	Core	Enhanced	Overall
Vox2Vox	0.900	0.813	0.757	0.824
[27]	(0.095)	(0.150)	(0.255)	(0.179)
Unetr	0.864	0.779	0.767	0.803
[12]	(0.132)	(0.166)	(0.249)	(0.248)
SwinUnet[1]	0.916	0.824	0.749	0.830
[11]	(0.056)	(0.137)	(0.252)	(0.169)
SwinUnet	0.848	0.731	0.633	0.737
[11]	(0.119)	(0.158)	(0.236)	(0.178)
nnUnet	0.911	0.825	0.755	0.830
[4]	(0.080)	(0.139)	(0.258)	(0.176)
TransBTS	0.857	0.762	0.656	0.759
[19]	(0.136)	(0.168)	(0.248)	(0.190)
HUT (Ours)	**0.921**	**0.834**	**0.792**	**0.849**
	(0.086)	(0.144)	(0.246)	(0.172)

It showed that a pure Transformer requires a large amount of data to achieve good results. This is why hybrid systems work better than the pure CNN and Transformer counterparts. For comparison, we include the Unetr method to illustrate the improvement of the accuracy of segmentation using the hybrid method. Unetr uses both Transformer and CNN architectures. Our evaluation shows a dice score of 0.781 and an HD95 score of 8.35 mm. That is an improvement over the SwinUnet without pre-training. On the contrary, our method, implemented based on a hybrid architecture, exceeds the dice score of the Unetr by 7.4% and HD95 by 42.7%. HUT also performs better than another hybrid architecture, TransBTS. It exceeds the dice score of TransBTS by 3.6% and HD95 by 23.4%.

Fig. 4 compares the performance of various methods for predicting segmentation, including Vox2Vox, Unetr, SwinUnet, nnUnet, and HUT on a representative slice taken from subject 49. Without pre-training, SwinUnet produces inaccurate results with visible artefacts. It can predict most edema but fails to accurately locate the tumour core and enhancing tumour. All methods can detect the presence of edema, but Unetr, a hybrid Unet and Transformer architecture has difficulty precisely predicting the tumour's location. nnUNet and Vox2Vox can identify the tumour core but not the enhancing tumour. HUT is the only method that accurately predicts all class labels.

In Fig. 5, nnUNet and Vox2Vox can identify the enhancing tumour but not the tumour core, while Unetr can locate the tumour core. Only HUT can accurately predict both the tumour core and the enhancing tumour. SwinUnet, conversely, fails to detect any tumour. In Fig. 6, SwinUnet again produces artefacts in the predicted results when there is no pre-training but performs well with pre-training. All other methods can accurately predict the presence of tumour and Edema.

3.4. Comparison of performances with existing methods on ischemic stroke lesion segmentation

Wong et al. [5] qualified anomalies less than 100 pixels as small lesions. Similarly, we also evaluated the performance of segmenting small lesions. In the experiment, we used the same criteria for the evaluation of the task of small lesion segmentation. The hyper-parameters of the models are again obtained through 5-fold cross-validation. As mentioned, the metrics used to evaluate the ischemic stroke lesion segmentation are Dice, HD95, IOU, Precision, and Recall. The mean and standard deviation of the metrics were computed from the test samples.

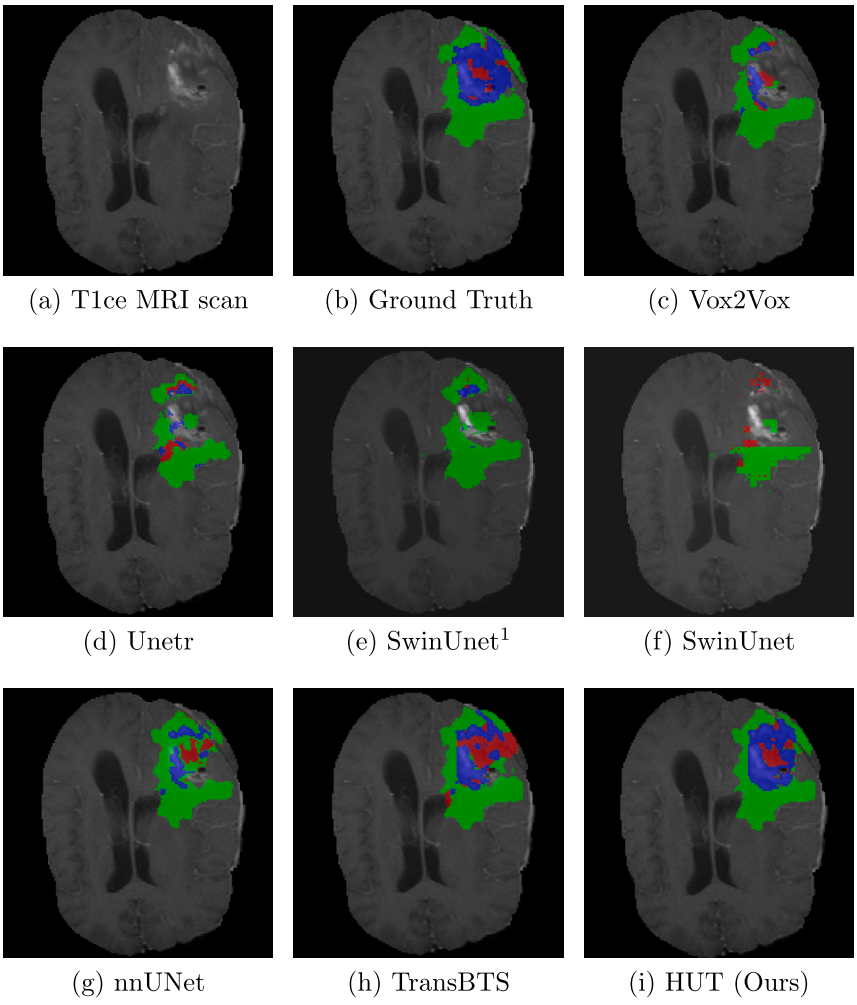


Fig. 4. Illustration of brain tumour segmentation of Subject 49 based on various methods.

As shown in Table 6, our HUT method improves the mean Dice score performance over the state-of-the-art SPiN [5] architecture by 4.84%. It also improves the mean 95th percentile Hausdorff Distance score (HD95) by 41%, as shown in the same table. It gains a dice score over DUNet by 34.5% and a dice score over X-Net by 17.5%. We found that nnUnet outperforms SPiN for all metrics except precision. This is likely due to the 3D CNN used in nnUnet compared to the sliding prediction of the 3D chunk used in SPiN. While Attn-Unet performs well, it falls short of nnUnet's performance. The incorporation of an attention mechanism results in a degradation of performance. HUT, however, can overcome this deficit using a Transformer network. It also performs significantly better than Unetr on the ATLASR12 dataset, with a 16.0% improvement in the dice score and a 55.2% improvement in HD95. All other methods described in [5] performed worse than HUT and are shown in Table 6. The same results for the task of small lesion segmentation are compared in Table 7, where HUT outperforms all other methods by a significant margin.

As observed in Table 7, HUT can achieve 3.60% and 18.6% improvement over the state-of-the-art SPiN [5] on the performance for the segmentation task on all and small lesions, respectively. It also improves the mean 95th percentile Hausdorff Distance score (HD95) by 44% and 42.6% for all and small lesions, respectively. Our method outperforms all other methods in [5] by a large margin for small lesion segmentation.

Fig. 7 compares the performance of various methods for predicting segmentation, including Unetr, SPiN, nnUnet, KiUnet, CLCInet, X-Net, Attn-Unet, and HUT. Most methods can accurately predict the location of a large lesion. Attention-based architectures like Attn-Unet and HUT tend to be more conservative in their predictions and therefore miss the "finger"-like part of the lesion. Fig. 8 shows a case where there is a very small lesion in the brain. Only HUT can locate the lesion, although Attn-Unet is close. Unetr detects the lesion accurately but also produces more artefacts. nnUnet, KiUnet and DUNet incorrectly detected the lesion beside the ventricle. The other methods fail to detect the lesion at all. Fig. 9 illustrates another case where HUT can correctly identify both lesion locations, although the prediction for the upper lesion is not as precise. nnUnet, CLCInet and Unetr can detect the upper lesion. Unetr correctly detects the larger lesion but fails to detect the other upper and smaller lesions. The remaining methods are unable to detect any lesion.

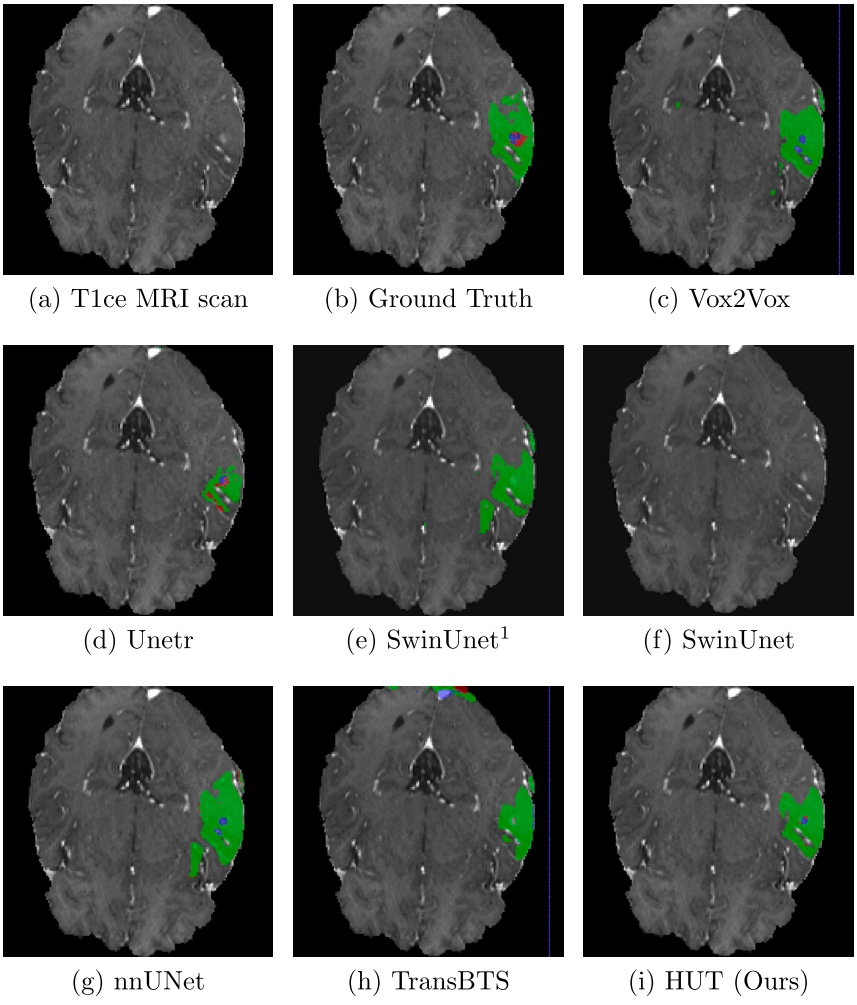


Fig. 5. Illustration of brain tumour segmentation of Subject 20 based on various methods.

Table 6
Comparison between mean and standard deviation (in parentheses) of Dice score and HD95 score (in mm) of the Ischemic Stroke lesion segmentation by HUT against SOTA methods.

Methods	Dice	HD95	IOU	Precision	Recall
nnUnet	0.713	14.294	0.568	0.767	**0.707**
[4]	(0.145)	(16.133)	(0.156)	(0.218)	(0.134)
Attn-Unet	0.698	16.803	0.553	0.751	0.685
[28]	(0.171)	(24.960)	(0.179)	(0.208)	(0.180)
Unetr	0.630	23.083	0.476	0.725	0.608
[12]	(0.148)	(22.046)	(0.152)	(0.767)	(0.176)
CLCI-Net	0.599	20.802	0.469	0.741	0.536
[29]	(0.257)	(22.644)	(0.232)	(0.258)	(0.276)
KiUnet	0.524	19.255	0.387	0.703	0.459
[30]	(0.226)	(16.290)	(0.206)	(0.237)	(0.241)
DUNet	0.548	22.809	0.404	0.652	0.521
[31]	(0.216)	(24.393)	(0.187)	(0.258)	(0.241)
X-Net	0.627	17.143	0.489	0.722	0.598
[32]	(0.216)	(15.897)	(0.204)	(0.208)	(0.264)
SPiN	0.703	17.427	0.556	0.806	0.654
[5]	(0.129)	(19.469)	(0.142)	(0.123)	(0.182)
HUT (ours)	**0.737**	**10.335**	**0.598**	**0.825**	0.706
	(0.127)	(10.074)	(0.144)	(0.172)	(0.153)

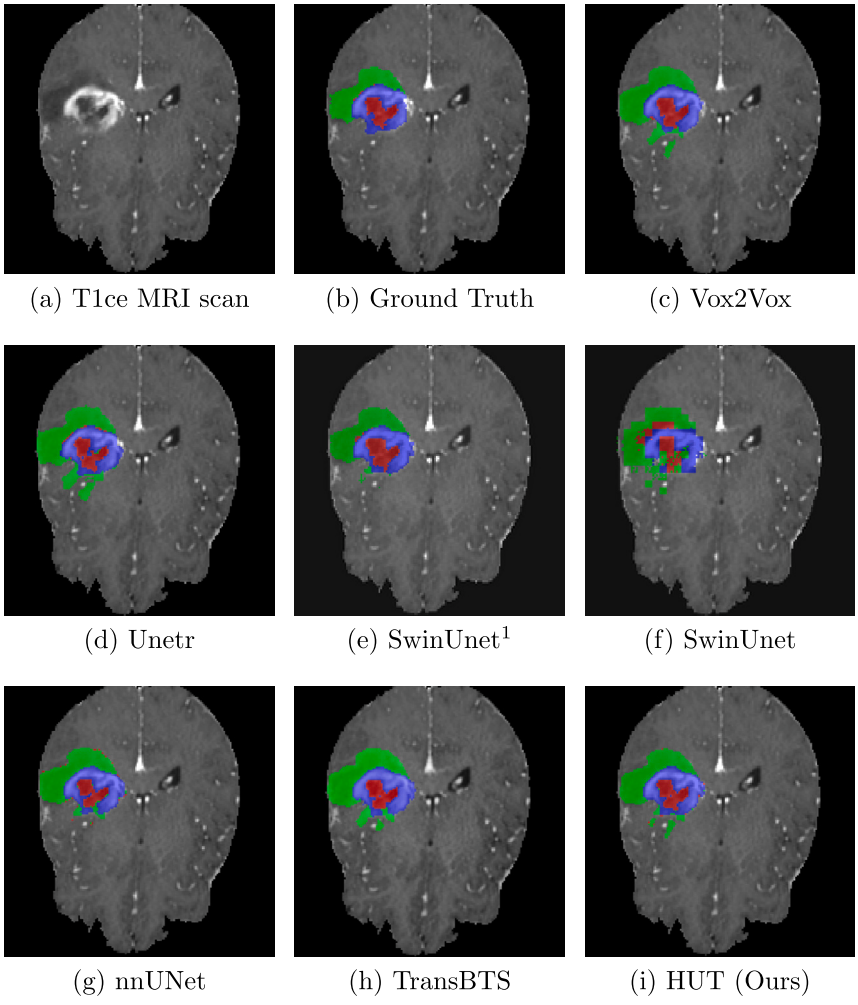


Fig. 6. Illustration of brain tumour segmentation of Subject 11 based on various methods.

Table 7
Comparison between mean and standard deviation (in parentheses) of Dice score and HD95 score (in mm) of the Ischemic Stroke small lesion segmentation by HUT against SOTA methods.

Methods	Dice	HD95	IOU	Precision	Recall
nnUnet	0.465	16.054	0.322	0.579	**0.515**
[4]	(0.190)	(12.081)	(0.168)	(0.291)	(0.219)
Attn-Unet	0.461	16.378	0.321	0.573	0.496
[28]	(0.194)	(16.946)	(0.175)	(0.277)	(0.218)
Unetr	0.385	19.710	0.258	0.580	0.375
[12]	(0.196)	(20.774)	(0.159)	(0.327)	(0.209)
CLCI-Net	0.246	22.884	0.178	0.417	0.215
[29]	(0.290)	(25.531)	(0.232)	(0.384)	(0.279)
KiUnet	0.246	15.979	0.173	0.466	0.206
[30]	(0.270)	(16.255)	(0.211)	(0.402)	(0.253)
DUNet	0.265	26.730	0.180	0.377	0.264
[31]	(0.250)	(23.336)	(0.188)	(0.332)	(0.269)
X-Net	0.335	22.885	0.237	0.491	0.309
[32]	(0.274)	(22.294)	(0.221)	(0.340)	(0.292)
SPiN	0.398	23.063	0.287	0.575	0.350
[5]	(0.274)	(20.764)	(0.229)	(0.332)	(0.272)
HUT (ours)	**0.472**	**12.630**	**0.327**	**0.634**	0.487
	(0.178)	(11.658)	(0.159)	(0.290)	(0.208)

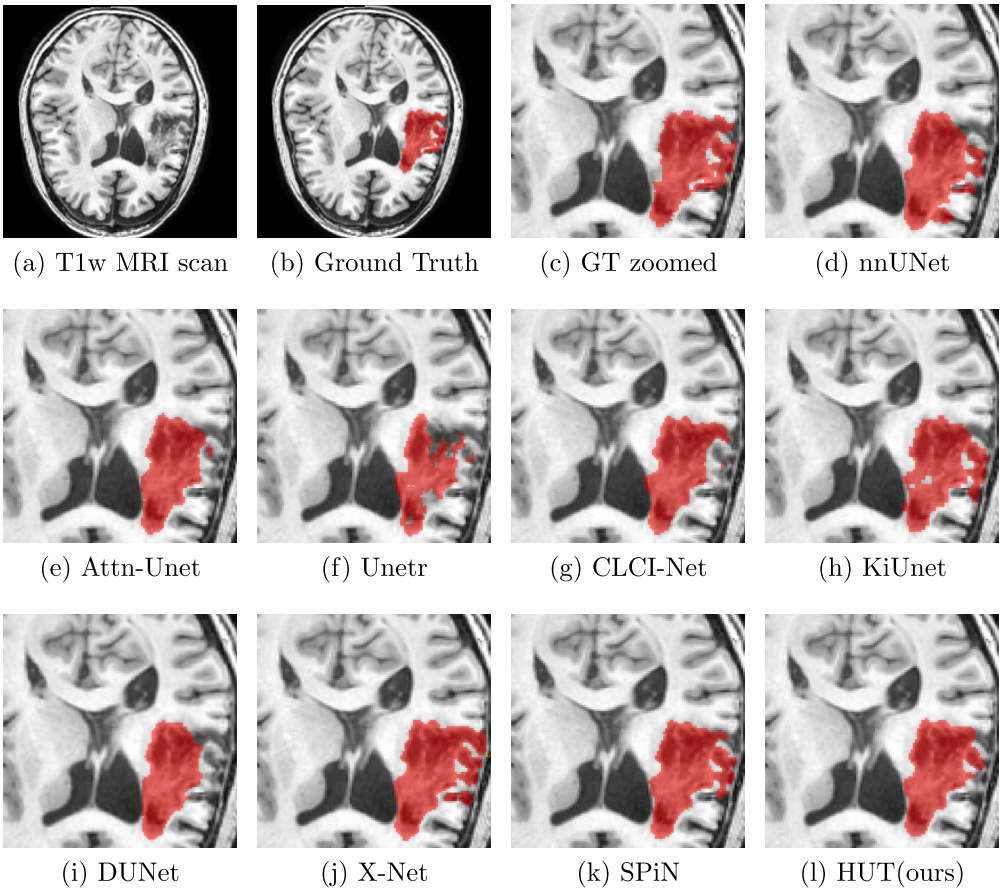


Fig. 7. Illustration of large lesion segmentation of Subject 08 based on various methods.

Table 8
Ablation study of mean Dice score and HD95 score (in mm) of the Brain Tumour segmentation (BraTS20) for HUT.

Methods	Dice	HD95	IOU	Precision	Recall
Without VTP	0.831	5.222	0.740	0.852	0.830
Baseline	0.836	6.043	0.745	0.846	0.846
Baseline + SS	0.839	5.739	0.749	0.859	**0.840**
Baseline + SS + Focal Loss	0.837	5.385	0.746	**0.869**	0.829
Baseline + SS + Dice Loss	**0.840**	**4.937**	**0.750**	0.861	0.839

Table 9
Ablation study of mean Dice score and HD95 score (in mm) of the Ischemic Stroke lesion segmentation (AtlasR12) for HUT.

Methods	Dice	HD95	IOU	Precision	Recall
Without VTP	0.713	14.294	0.568	0.767	0.707
Baseline	0.720	13.639	0.579	0.785	0.700
Baseline + SS	0.734	10.465	**0.601**	0.698	**0.801**
Baseline + SS + Focal Loss	0.732	11.175	0.601	0.795	0.698
Baseline + SS + Dice Loss	0.699	12.935	0.557	0.782	0.684
Baseline + SS + Balancing	**0.737**	**10.335**	0.598	**0.825**	0.706

3.5. Ablation studies

We performed ablation studies of our method over the BraTS20 and ATLAS dataset. As shown in Tables 8 and 9, we compare the performance of the baseline method at various stages. The baseline method uses only cross-entropy loss as the objective loss function.

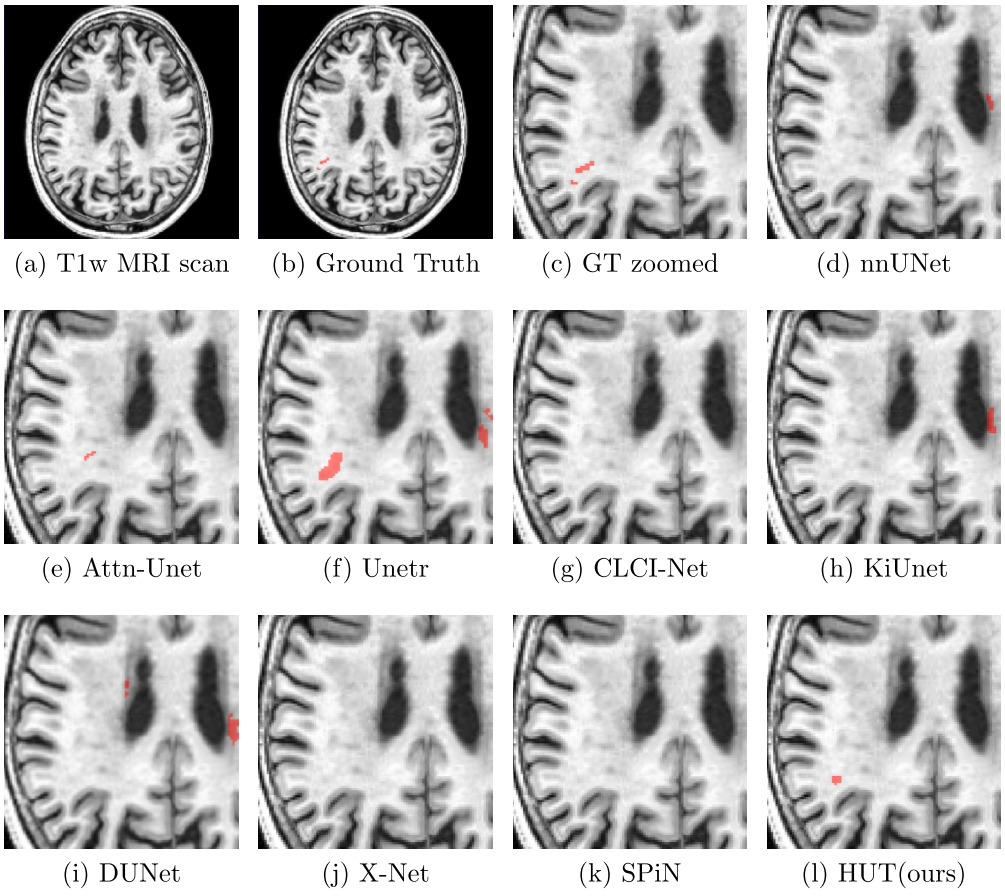


Fig. 8. Illustration of small lesion segmentation of Subject 26 based on various methods. Only HUT captured the correct location of the lesion.

It does not include self-supervised (SS) CLS training at the output of the cross-transformer projection header of the CLS tokens by default.

As observed in Table 8, the baseline model gains a dice score of 0.62% over the architecture without the VTP, on the BraTS20 dataset. The best-performing model, which includes the Dice and self-supervised losses as part of the training, gains a dice score of 0.48% over the baseline. As for the ablation study on ATLASR12 dataset shown in Table 9, the proposed baseline model performs with a dice score of 0.720 and an HD95 of 13.64 mm, which still performs better than SPiN. The baseline model gains a dice score of 0.98% over the architecture without the VTP.

Adding soft dice loss and self-supervised (SS) CLS training to the baseline causes a decline in performance. Soft dice loss [33] is a loss function that alleviates the class imbalance issue by appropriately computing the difference between unity and the dice score. With focal loss and SS, the dice score improves to 0.732. A focal loss [34] is another loss function which tries to address the issue of a class imbalance in segmentation. The focal loss function down-weight the loss contributed by the easy examples by a modulating factor. Therefore, the loss for the harder examples will be relatively higher.

On the contrary, the model performs optimally using cross-entropy loss with a weighting of 0.15 for the background and 0.85 for the foreground. The model performs slightly worse than optimal without this weighting or balancing component. Weighting [3] also mainly addresses the imbalance issue of the datasets as, in most cases, the portion of the background dominates the amount of the class label (lesion). It exerts more emphasis on the class label rather than the background. Therefore, the proposed architecture of HUT is best trained using the weighted cross-entropy loss function in all of these examples to address the class imbalance problem. We note that the class imbalance issue is closely related to the ability to detect a very small lesion in these ablation studies.

4. Discussion and conclusion

We proposed a deep neural network architecture called HUT for brain lesion and tumour segmentation. Our method showed equal or better performance in segmenting tumours and lesions over existing state-of-the-art methods. We use a hybrid of UNet and a cross-resolution transformer to segment various brain anomalies. The cross-resolution transformer generates two different resolutions combined with UNet's skip connections. We found that using two transformers, one for small patches and another for

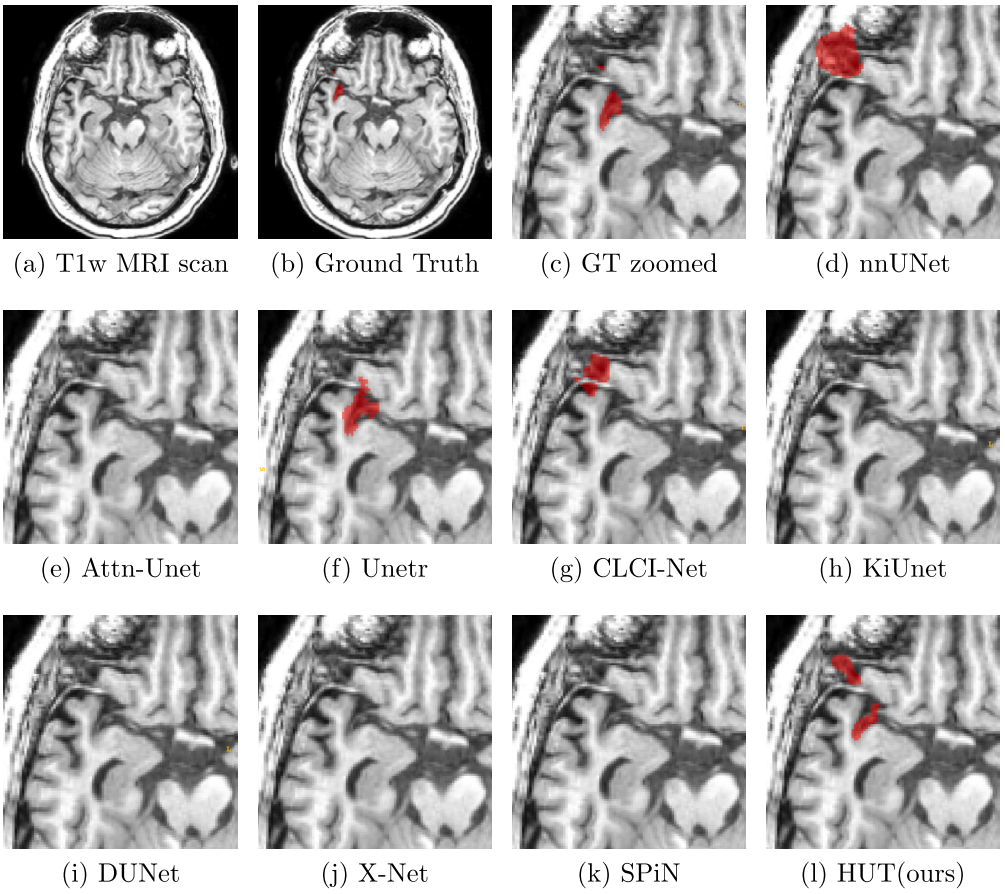


Fig. 9. Illustration of small lesion segmentation of Subject 05 based on various methods. Only HUT captured the correct location of the lesion.

larger patches, followed by the cross transformer, helps improve performance. We also employ CLS tokens to generate attention maps and classification headers, which help improve convergence and performance.

However, we have observed that a pure Transformer-based architecture performs poorly in segmentation tasks without much training data. In [11], the authors used a pre-trained Swin Transformer on the Imagenet dataset [35] and achieved a dice score. Our proposed hybrid UNet Transformer model improved the dice score and HD95 without pre-training. The advantage of the architecture is that it is more data-efficient, and the training converges much faster than the pure Transformer architecture.

In addition, HUT is designed to address the local and long-range correlations between voxels and so outperformed current methods using Transformer, UNet, and GAN architectures for medical image segmentation by a significant margin. We also implemented a self-supervised strategy to improve performance, significantly improving the Dice score and Hausdorff distance. These results have important implications for medical imaging applications. Despite the gain in overall performance, HUT has a lower recall but a higher precision in lesion segmentation. It implies the under-segmentation of the technique on the ATLASR12 dataset, which has more small anomalies to segment. Future work will look into enhancing the architecture to address this drawback.

Additional information

No additional information is available for this paper.

CRediT authorship contribution statement

Wei Kwek Soh: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Writing – original draft, Writing – review & editing. **Hing Yee Yuen:** Data curation, Formal analysis. **Jagath C. Rajapakse:** Conceptualization, Investigation, Methodology, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data associated with this study has been deposited at https://www.med.upenn.edu/cbica/brats2020/data.html and https://fcon_1000.projects.nitrc.org/indi/retro/atlas.html.

References

[1] C. Board, Cancer net, American Society of Clinical Oncology (ASCO), 2022, Online. Available: https://www.cancer.net/cancer-types/brain-tumor/statistics (Accessed 2022).
[2] C.W. Tsao, A.W. Aday, Z.I. Almarzooq, A. Alonso, A.Z. Beaton, M.S. Bittencourt, A.K. Boehme, A.E. Buxton, A.P. Carson, Y. Commodore-Mensah, et al., Heart disease and stroke statistics—2022 update: a report from the American heart association, Circulation 145 (8) (2022) e153–e639.
[3] O. Ronneberger, P. Fischer, T. Brox, U-net: convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2015, pp. 234–241.
[4] F. Isensee, P.F. Jaeger, S.A. Kohl, J. Petersen, K.H. Maier-Hein, nnu-net: a self-configuring method for deep learning-based biomedical image segmentation, Nat. Methods 18 (2) (2021) 203–211.
[5] A. Wong, A. Chen, Y. Wu, S. Cicek, A. Tiard, B.-W. Hong, S. Soatto, Small lesion segmentation in brain mris with subpixel embedding, in: International MICCAI Brainlesion Workshop, Springer, 2022, pp. 75–87.
[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, arXiv preprint, arXiv:1706.03762, 2017.
[7] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: pre-training of deep bidirectional transformers for language understanding, arXiv preprint, arXiv:1810.04805, 2018.
[8] B. Gheflati, H. Rivaz, Vision transformer for classification of breast ultrasound images, arXiv preprint, arXiv:2110.14731, 2021.
[9] J.M.J. Valanarasu, P. Oza, I. Hacihaliloglu, V.M. Patel, Medical transformer: gated axial-attention for medical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2021, pp. 36–46.
[10] Y. Xie, J. Zhang, C. Shen, Y. Xia, Cotr: efficiently bridging cnn and transformer for 3d medical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2021, pp. 171–180.
[11] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, M. Wang, Swin-unet: unet-like pure transformer for medical image segmentation, arXiv preprint, arXiv:2105.05537, 2021.
[12] A. Hatamizad, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H.R. Roth, D. Xu, Unetr: transformers for 3d medical image segmentation, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022, pp. 574–584.
[13] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A.L. Yuille, Y. Zhou, Transunet: transformers make strong encoders for medical image segmentation, arXiv preprint, arXiv:2102.04306, 2021.
[14] Y. Gu, Z. Piao, S.J. Yoo, Sthardnet: swin transformer with hardnet for mri segmentation, Appl. Sci. 12 (1) (2022) 468.
[15] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10012–10022.
[16] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16×16 words: transformers for image recognition at scale, arXiv preprint, arXiv:2010.11929, 2020.
[17] Y. Tang, D. Yang, W. Li, H. Roth, B. Landman, D. Xu, V. Nath, A. Hatamizadeh, Self-supervised pre-training of swin transformers for 3d medical image analysis, arXiv preprint, arXiv:2111.14791, 2021.
[18] H. Wang, S. Xie, L. Lin, Y. Iwamoto, X.-H. Han, Y.-W. Chen, R. Tong, Mixed transformer u-net for medical image segmentation, arXiv preprint, arXiv:2111.04734, 2021.
[19] W. Wang, C. Chen, M. Ding, H. Yu, S. Zha, J. Li, Transbts: multimodal brain tumor segmentation using transformer, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2021, pp. 109–119.
[20] B.H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, et al., The multimodal brain tumor image segmentation benchmark (brats), IEEE Trans. Med. Imaging 34 (10) (2014) 1993–2024.
[21] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J.S. Kirby, J.B. Freymann, K. Farahani, C. Davatzikos, Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features, Sci. Data 4 (1) (2017) 1–13.
[22] S. Bakas, M. Reyes, A. Jakab, S. Bauer, M. Rempfler, A. Crimi, R.T. Shinohara, C. Berger, S.M. Ha, M. Rozycki, 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.
[23] C.-F.R. Chen, Q. Fan, R. Panda, Crossvit: cross-attention multi-scale vision transformer for image classification, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 357–366.
[24] W. Wang, L. Yao, L. Chen, B. Lin, D. Cai, X. He, W. Liu, Crossformer: a versatile vision transformer hinging on cross-scale attention, arXiv preprint, arXiv:2108.00154, 2021.
[25] S.-L. Liew, J.M. Anglin, N.W. Banks, A large, open source dataset of stroke anatomical brain images and manual lesion, Sci. Data 5 (1) (2018).
[26] S.-L. Liew, The Anatomical Tracings of Lesions after Stroke (ATLAS) Dataset-Release 1.2, 2018, 2017.
[27] M.D. Cirillo, D. Abramian, A. Eklund, Vox2vox: 3d-gan for brain tumour segmentation, in: International MICCAI Brainlesion Workshop, Springer, 2020, pp. 274–284.
[28] M. Islam, V. Vibashan, V. Jose, N. Wijethilake, U. Utkarsh, H. Ren, Brain tumor segmentation and survival prediction using 3d attention unet, in: International MICCAI Brainlesion Workshop, Springer, 2020, pp. 262–272.
[29] H. Yang, W. Huang, K. Qi, C. Li, X. Liu, M. Wang, H. Zheng, S. Wang, Clci-net: cross-level fusion and context inference networks for lesion segmentation of chronic stroke, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2019, pp. 266–274.
[30] J.M.J. Valanarasu, V.A. Sindagi, I. Hacihaliloglu, V.M. Patel, Kiu-net: towards accurate segmentation of biomedical images using over-complete representations, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2020, pp. 363–373.
[31] Q. Jin, Z. Meng, T.D. Pham, Q. Chen, L. Wei, R. Su, Dunet: a deformable network for retinal vessel segmentation, Knowl.-Based Syst. 178 (2019) 149–162.
[32] K. Qi, H. Yang, C. Li, Z. Liu, M. Wang, Q. Liu, S. Wang, X-net: brain stroke lesion segmentation based on depthwise separable convolution and long-range dependencies, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2019, pp. 247–255.
[33] F. Milletari, N. Navab, S.-A. Ahmadi, V-net: fully convolutional neural networks for volumetric medical image segmentation, in: 2016 Fourth International Conference on 3D Vision (3DV), IEEE, 2016, pp. 565–571.
[34] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2980–2988.
[35] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: a large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2009, pp. 248–255.