A Volumetric Transformer for Accurate 3D Tumor Segmentation

Himashi Peiris
*Department of Electrical and Computer Systems Engineering,*
*Faculty of Engineering,*
*Monash University,*
*Melbourne, Australia.

Munawar Hayat
*Department of Data Science & AI,*
*Faculty of IT,*
*Monash University,*
*Melbourne, Australia.

Zhaolin Chen
*Monash Biomedical Imaging (MBI),*
*Monash University,*
*Melbourne, Australia.

Gary Egan
*Monash Biomedical Imaging (MBI),*
*Monash University,*
*Melbourne, Australia.

Mehrtash Harandi
*Department of Electrical and Computer Systems Engineering,*
*Faculty of Engineering,*
*Monash University,*
*Melbourne, Australia.

Abstract

This paper presents a Transformer architecture for volumetric medical image segmentation. Designing a computationally efficient Transformer architecture for volumetric segmentation is a challenging task. It requires keeping a complex balance in encoding local and global spatial cues, and preserving information along all axes of the volumetric data. The proposed volumetric Transformer has a U-shaped encoder-decoder design that processes the input voxels in their entirety. Our encoder has two consecutive self-attention layers to simultaneously encode local and global cues, and our decoder has novel parallel shifted window based self and cross attention blocks to capture fine details for boundary refinement by subsuming Fourier position encoding. Our proposed design choices result in a computationally efficient architecture, which demonstrates promising results on Brain Tumor Segmentation (BraTS) 2021, and Medical Segmentation Decathlon (Pancreas and Liver) datasets for tumor segmentation. We further show that the representations learned by our model transfer better across-datasets and are robust against data corruptions. Our code implementation is publicly available.

Keywords: Deep Learning, Volumetric Transformer, 3D Medical Image Segmentation
Figure 1: The model size vs Dice Similarity Score is shown in this plot. Circle size indicates Computational Complexity by FLOPs. The proposed method (VT-UNet) achieves highest Dice Similarity Score compared to baselines/SOTA methods while maintaining a smaller model size and low computational complexity.

1. Introduction

In this paper, we propose a Transformer Vaswani et al. (2017) based encoder-decoder architecture, which can directly process 3D volumetric data (instead of dividing it into 2D slices) for the task of volumetric semantic segmentation. Transformer-based Vaswani et al. (2017) architectures have recently gained growing success in computer vision tasks Khan et al. (2021). Compared with their CNN counterparts, models based on the transformer architecture not only achieve better empirical performances Dosovitskiy et al. (2020); Touvron et al. (2021); Strudel et al. (2021); Carion et al. (2020); Zhu et al. (2020); Girdhar et al. (2019); Arnab et al. (2021); Liu et al. (2021b), but have demonstrated better robustness against data corruptions and occlusions Naseer et al. (2021), high frequency image noise Shao et al. (2021), and adversarial perturbations Shao et al. (2021); Naseer et al. (2021); Paul and Chen (2021); Mao et al. (2021); Bhojanapalli et al. (2021). Furthermore, it is argued that transformers are a promising candidate towards a unified architecture across different data modalities and tasks Jaegle et al. (2021b,a).

Inspired by the strong empirical results of the transformer based models on vision tasks including image classification Dosovitskiy et al. (2020); Touvron et al. (2021), object detection Carion et al. (2020); Zhu et al. (2020), video recognition Girdhar et al. (2019); Arnab et al. (2021); Liu et al. (2021b), and semantic segmentation Ye et al. (2019), their promising generalization and robustness characteristics Shao et al. (2021); Naseer et al. (2021), and their flexibility to model long range interactions, we propose a volumetric transformer architecture for segmenting 3D medical image modalities (e.g., MRI, CT), called VT-UNet. Earlier efforts to develop transformer based segmentation models for 3D medical images,
such as the work of Cao et al. (2021), have been shown to outperform state-of-the-art CNN counterparts. However, these methods opt for slicing 3D volumes into 2D images and process 2D images as inputs (Cao et al. (2021); Chen et al. (2021b,a)). As such, considerable and potentially critical volumetric information, essential to encapsulating inter-slice dependencies, is lost. While some hybrid approaches (using both convolutional blocks and Transformer layers) keep the 3D volumetric data intact (Wang et al. (2021); Hatamizadeh et al. (2021); Zhou et al. (2021)), the design of transformer based architecture, capable of keeping intact the volumetric data at input, is yet unexplored in the literature. Our work takes the first step in this direction, and proposes a model, which not only achieves better segmentation performance, but also demonstrates better robustness against data artefacts. Further, our model is efficient in terms of number of parameters and FLOPS (see Fig. 1), and its pre-trained features generalize better across other datasets.

The design of our Transformer based volumetric segmentation model is built upon the seminal encoder-decoder based UNet architecture (Ronneberger et al. (2015)), which has gained significant traction not only for segmentation task (Ronneberger et al. (2015); Zhou et al. (2018); Oktay et al. (2018); Zhang et al. (2020); Xia et al. (2020), but also for reconstruction (Hyun et al. (2018); Han et al. (2019), denoising (Zamir et al. (2020); Wu et al. (2019), and super-resolution (Anwar et al. (2020); Wang et al. (2018) to name a few.

While the Transformer models have highly dynamic and flexible receptive field and do an excellent job in capturing long-range interactions, designing a Transformer based UNet architecture for volumetric segmentation is a challenging task. This is because: (1) Encapsulating voxel information and capturing the connections between arbitrary positions in the volumetric sequence is not straightforward. Compared with Transformer based approaches for 2D image segmentation (Cao et al. (2021)), the data in each slice of the volume is connected to three views and discarding either of them can be detrimental. (2) Preserving spatial information in the volumetric data is a daunting task. Even for 2D images, while breaking the image into patches and projecting patches into tokens as introduced in Vision Transformer (ViT), local structural cues can be lost, as shown in Tokens-to-token ViT (Yuan et al. (2021)). Effectively encoding the local cues while simultaneously capturing global interactions along multiple axes of the volumetric data is therefore a challenging task. (3) Due to the quadratic complexity of the self-attention, and large sized 3D volume tensor inputs, designing a Transformer based segmentation model, which is computationally efficient, requires careful design considerations.

Our proposed VT-UNet model effectively tackles the above design challenges by proposing a number of modules. In our UNet based architecture, we develop two types of Transformer blocks. First, our blocks in the encoder which directly work on the 3D volumes, in a hierarchical manner, to jointly capture the local and global information, similar in spirit to the Swin Transformer blocks (Liu et al. (2021a)). Secondly, for the decoder, we introduce parallel cross-attention and self-attention in the expansive path, which creates a bridge between queries from the decoder and keys & values from the encoder. By this parallelization of the cross-attention and self-attention, we aim to preserve nearly the full global context during the decoding process, which is important for the task of segmentation. Our parallelization is then coupled with the sinusoidal version of the Fourier feature positional encoding to further improve the learning capability of the resulting network. Since VT-UNet is free from convolutions and combines attention outputs from two modules during the decoding,
the order of the sequence is important to get accurate predictions. Inspired by Vaswani et al. (2017), apart from applying relative positional encoding while computing attention in each Transformer block, we augment the decoding process and inject the complementary information extracted from Fourier feature positions of the tokens in the sequence.

In summary, our major contributions are, (1) We reformulate volumetric tumor segmentation from a sequence-to-sequence perspective, and propose a UNet shaped Volumetric Transformer for multi-modal medical image segmentation. (2) We design an encoder block with two consecutive self attention layers to jointly capture local and global contextual cues. Further, we design a decoder block which enables parallel (shifted) window based self and cross attention. This parallelization uses one shared projection of the queries and independently computes cross and self attention. To further enhance our features in the decoding, we propose a convex combination approach along with Fourier positional encoding. (3) Incorporating our proposed design choices, we substantially limit the model parameters while maintaining lower FLOPs compared to existing approaches (e.g., 3D UNet Çiçek et al. (2016), UNETR Hatamizadeh et al. (2021)). (4) We conduct extensive evaluations and show that our design consistently achieves state-of-the-art volumetric segmentation results, along with enhanced robustness to data artefacts, and better cross-dataset generalization of the pretrained features.

2. Methodology

![Figure 2: (a) Illustrates the Proposed VT-UNet Architecture. (b) shows Volumetric Transformer (VT) Encoder-Decoder Structure. Here, $k$ denotes number of classes segmented using VT-UNet.](image)

We denote vectors and matrices in bold lower-case $\mathbf{x}$ and bold upper-case $\mathbf{X}$, respectively. Norm of a vector is denoted by $\| \cdot \|$ and $\| \mathbf{x} \|_1 = \sum_i |[x_i]|$. The inner product between vectors is shown by $\langle \cdot, \cdot \rangle$ and $\| \mathbf{x} \|_2 = \langle \mathbf{x}, \mathbf{x} \rangle$. When norms and inner product are used over
3D tensors, we assume tensors are flattened accordingly. For example, for 3D tensors $\mathbf{A}$ and $\mathbf{B}$, $\langle \mathbf{A}, \mathbf{B} \rangle = \sum_{i,j,k} A[i,j,k]B[i,j,k]$ and $\|\mathbf{A}\|_1 = \sum_{i,j,k} |A[i,j,k]|$.

**Preliminaries.** We start by providing a brief description about the Self-Attention (SA) mechanism that empowers Transformers Vaswani et al. (2017) (see Khan et al. (2021) for an in-depth discussion).

Let $\mathbf{X} = \{x_1, x_2, \ldots, x_T\}, x_i \in \mathbb{R}^C$ be a sequence representing a signal of interest (e.g., an MRI volume). We call each $x_i$ a token. In SA, we are interested in generating a new sequence of tokens, $\{z_1, z_2, \ldots, z_T\}$, from $\mathbf{X}$ to better represent our signal according to the objective of learning. In doing so, we can assume that $x_i$s span a subset of $\mathbb{R}^C$ and define $z_i$ as a point in that span, i.e. $z_i = \sum_j a_{ij}x_j$, where $a_{ij}$s are combination values defined by $x_i$ and $x_j$. A possible choice for $a_{ij}$ is based on the similarity of $x_i$ and $x_j$, which algebraically is proportional to $\langle x_i, x_j \rangle$. Such a choice enables us to define the token $z_i$ by attending to important parts of the input sequence according to the objective in hand, hence the name attention. We can take a further step and put a constraint on our design by enforcing the generated tokens to lie inside the convex hull defined by $x_is$. In that case, we will have $z_i = \sum_j a_{ij}x_j, a_{ij} \geq 0, \sum_j a_{ij} = 1$. The convex hull formulation endows nice properties, one being that the resulting tokens cannot grow boundlessly, given the fact that the input is assumed to be a natural signal.

The SA operation is built upon the above idea with some modifications. Firstly, we assume that tokens, in their original form, might not be optimal for defining the span. Therefore, in SA we define the span by learning a linear mapping from the input tokens. This we can show with $\mathbb{R}^{T \times C_x} \ni \mathbf{V} = \mathbf{XW}_V$, where we stack tokens $x_i$s into the rows of $\mathbf{X}$ (i.e., $\mathbf{X} = [x_1 | x_2 | \cdots | x_T]^{\top}$). Then we turn our attention to $a_{ij}$ and define it by learning two linear mappings, following a similar argument. In particular, first we define a set of keys from $x_i$ as $\mathbb{R}^{T \times C_k} \ni \mathbf{K} = \mathbf{XW}_K$. To generate $a_{ij}$, we measure the similarity of a transformed version of $x_i$, which we call the query $q_i = \mathbf{W}_Qx_i$, with respect to the keys $k_j$ that are represented by the rows of $\mathbf{K}$. That is, $a_{ij} \propto \langle q_i, k_j \rangle$. Put everything into a matrix form and opt for a softmax function to achieve the constraints $a_{ij} \geq 0$ and $\sum_j a_{ij} = 1$, we arrive at

$$\text{SA}(Q, K, V) = \text{SoftMax} \left( \frac{QK^\top}{\sqrt{C}} \right) V,$$

with $V = \mathbf{XW}_V$, $K = \mathbf{XW}_K$, and $Q = \mathbf{XW}_Q$.

Following previous work by Hu et al. Hu et al. (2019), we use a slight modification of the self-attention in our task as follows:

$$\text{SA}(Q, K, V) = \text{SoftMax} \left( \frac{QK^\top}{\sqrt{C}} + B \right) V,$$

where $\mathbb{R}^{r \times r} \ni B$ is trainable and acts as a relative positional bias across tokens in the volume. In practice, computing SA for multiple attention heads several times in parallel is called Multi-head Self-Attention (MSA). Eq. (2) is the basic building block of our Volumetric Transformer Window based Multi-head Self-Attention (VT-W-MSA), and the Volumetric Transformer Shifted Window based Multi-head Self-Attention (VT-SW-MSA), discussed next.
Overview of VT-UNet. Fig. 2 shows the conceptual diagram of the proposed volumetric transformer network or VT-UNet for short. The input to our model is a 3D volume of size $D \times H \times W \times C$. The output is a $D \times H \times W \times K$ dimensional volume, representing the presence/absence of voxel-level class labels ($K$ is the number of classes). The main blocks in VT-UNet are: 1. 3D patch partitioning, 2. VT encoder block, 3. 3D patch merging, 4. VT decoder block, and 5. 3D patch expanding. Below, we discuss architectural form of the VT-UNet modules and explain the functionality and rationals behind our design in detail.

2.1 The VT Encoder

In this section we discuss about individual modules of VT encoder which consists of 3D Patch Partitioning layer together with Linear Embedding layer, and 3D Patch merging layer followed by two successive VT encoder blocks.

3D Patch Partitioning. Transformer-based models work with a sequence of tokens. The very first block of VT-UNet accepts a $D \times H \times W \times C$ dimensional medical volume (e.g., MRI) and creates a set of tokens by splitting the 3D volume into non-overlapping 3D patches (see Fig. 3). The size of partitioning kernel is $P \times M \times M$, resulting in describing the volume by $\tau = \lfloor D/P \rfloor \times \lfloor H/M \rfloor \times \lfloor W/M \rfloor$ tokens. The 3D patch partitioning is followed by a linear embedding to map each token with dimensionality $P \times M \times M$ to a $C$ dimensional vector. Typical values for $M$, $P$ and $C$ according to our experiments are 4, 4, and 96, respectively.

VT Encoder Block. In ViT, tokens carry significant spatial information due to the way they are constructed. The importance of performing SA by windowing in ViT has been shown in several recent studies, most notably in Swin Transformer Liu et al. (2021a). Following a similar principal in the design of Swin Transformers, albeit for volumetric data, we propose 3D windowing operations in our VT Encoder Blocks (VT-Enc-Blks). In particular, we propose two types of windowing, namely regular window and shifted window, which we show by VT-W-MSA and VT-SW-MSA for simplicity, respectively. Fig. 2b provides the design specifics of VT-W-MSA and VT-SW-MSA, while Fig. 3 illustrates the windowing operation. Both VT-W-MSA and VT-SW-MSA employ attention layers with windowing, followed by a 2-layer Multi Layer Perceptron (MLP) with Gaussian Error Linear Unit (GELU) Hendrycks and Gimpel (2016) non-linearity in between. A Layer Normalization (LN) is applied before every MSA and MLP, and a residual connection is applied after each module.

The windowing enables us to inject inductive bias in modeling long range dependencies between tokens. In both VT-W-MSA and VT-SW-MSA, attention across tokens within a window helps representation learning. In the VT-W-MSA, we split the volume evenly into smaller non-overlapping windows as illustrated in Fig. 3. Since tokens in adjacent windows cannot see each other with VT-W-MSA, we make use of a shifted window in VT-SW-MSA (see the right most panel Fig. 3) which bridges tokens in adjacent windows of VT-W-MSA. The windowing is inspired by the Swin Transformer Liu et al. (2021a) and can be understood as generalization to volumetric data.\footnote{While preparing our draft, we came across Video Transformer by Liu et al. Liu et al. (2021b) where Swin-Transformer was extended to video modality. The windowing operation in our work resembles the operation done in the video Transformer, that extends the benefits of windowing beyond images.}
Figure 3: Visualization of Volumetric Shifted Windows. Assume that the MRI volume is $D \times H \times W$ is $8 \times 8 \times 8$ and the window size for partitioning the volume $P \times M \times M$ is $4 \times 4 \times 4$. Here, layer $l$ adopts the regular window partition in the first step of VT block which results in $2 \times 2 \times 2 = 8$ number of windows. Inside layer $l+1$, volumetric windows are shifted by $(P/2, M/2, M/2) = (2, 2, 2)$ tokens. This results in $3 \times 3 \times 3 = 27$ number of windows.

the following functionality:

$$
\hat{z}^l = \text{VT-W-MSA} \left( \text{LN} \left( z^{l-1} \right) \right) + z^{l-1}, \\
z^l = \text{MLP} \left( \text{LN} \left( \hat{z}^l \right) \right) + \hat{z}^l, \\
\hat{z}^{l+1} = \text{VT-SW-MSA} \left( \text{LN} \left( z^l \right) \right) + z^l, \\
z^{l+1} = \text{MLP} \left( \text{LN} \left( \hat{z}^{l+1} \right) \right) + \hat{z}^{l+1},
$$

(3)

where $\hat{z}^l$ and $z^l$ denote the output features of the VT-W-MSA module and the MLP module for block $l$, respectively.

A note on computational complexity. The computational complexity of the SA described in Eq. (2) is dictated by computations required for obtaining $Q, K, V$, computing $QK^\top$ and obtaining the resulting tokens by applying the output of the Softmax (which is a $\tau \times \tau$ matrix) to $V$. This adds up to $\mathcal{O}(3\tau C^2 + 2\tau^2 C)$, where $C$ and $\tau$ are the dimensionality and the number of tokens, respectively. Windowing will reduce the computational load of the SA according to $\mathcal{O}(3\tau C^2 + 2\tau \kappa C)$ where we have assumed that tokens are grouped into $\kappa$ windows and SA is applied within each window. In our problem, where tokens are generated from volumetric data, $\tau \gg \kappa$ and hence windowing not only helps in having better discriminatory power, but also it helps in reducing the computational load of the algorithm$^2$.

2. For the sake of simplicity and explaining the key message, we have made several assumptions in our derivation. First, we have assumed $C_k = C_v = C$. We also did not include the FLOPs needed to compute the softmax. Also, in practice, one uses a multi-head SA, where the computation is break down across several parallel head working on lower dimensional spaces (e.g., on for $V$, we use $C/h$ dimensional spaces where $h$ is the number of heads). This will reduce the computational load accordingly. That said, the general conclusion provided here is valid.
3D Patch Merging. We make use of 3D patch merging blocks to generate feature hierarchies in the encoder of VT-UNet. Having such hierarchies is essential to generate finer details in the output for the dense prediction tasks Liu et al. (2021a); Chen et al. (2017).

After every VT-Enc-Blk, we merge adjacent tokens along the spatial axes in a non-overlapping manner to produce new tokens. In doing so, we first concatenate features of each group of $2 \times 2$ neighboring tokens. The resulting vector is projected via a linear mapping to a space where the channel dimensionality of the tokens is doubled (see Fig. 2).

The benefit of patch merging is not limited to feature hierarchies. The computational complexity of SA is quadratic in the number of tokens Liu et al. (2021a,b). As such, patch merging reduces the FLOPs count of the VT-UNet by a factor of 16 after each VT-Enc-Blk. To give the reader a better idea and as we will discuss in §Sec. 4, the tiny VT-UNet model uses only 6.7% FLOPs in comparison to its fully volumetric CNN counterpart Milletari et al. (2016) while achieving a similar performance (slightly better indeed)! Please note that the patch merging block is not used in the bottleneck stage.

2.2 The VT Decoder

After bottleneck layer which consists of a VT-Enc-Blk together with 3D Patch Expanding layer, the VT decoder starts with successive VT Decoder Blocks (VT-Dec-Blks), 3D patch expanding layers and a classifier at the end to produce the final predictions. There are some fundamental design differences between VT-Enc-Blk and VT-Dec-Blk which we will discuss next.

3D Patch Expanding. The functionality of patch expanding is to somehow revert the effect of patch merging. In other words and in order to construct the output with the same spatial-resolution as the input, we need to create new tokens in the decoder. For the sake of discussion, consider the patch expanding after the bottleneck layer (see the middle part of Fig. 2). The input tokens to the patch expanding are of dimensionality $8C$. In the patch expanding, we first increase the dimensionality of the input tokens by a factor of two using a linear mapping. Following a reshaping, we can obtain $2 \times 2$ tokens with dimensionality $4C$ from the resulting vector of dimensionality $2 \times 8C$. This, we will reshape along the spatial axes and hence for $D/4 \times H/32 \times W/32 \times 8C$, we create $D/4 \times H/16 \times W/16 \times 4C$ tokens.

VT Decoder Block. The UNet Ronneberger et al. (2015) and its variants Oktay et al. (2018); Zhou et al. (2018) make use of lateral connections between the encoder and the decoder to produce fine-detailed predictions. This is because the spatial information is lost, at the expense of attaining higher levels of semantics, as the input passes through the encoder. The lateral connections in the UNet makes it possible to have the best of both worlds, spatial information from lower layers and semantic information from upper layers (along the computational graph). Having this in mind, we propose a hybrid form of SA at the decoder side (see Fig. 2b for an illustration). Each VT-Dec-Blk receives the generated tokens of its previous VT-Dec-Blk along with the key ($K_E$) and value ($V_E$) tokens from the VT-Enc-Blk sitting at the same stage of VT-UNet, see Fig. 2a. Recall that a VT-Enc-Blk has two SA blocks with regular and shifted windowing operations. VT-Dec-Blk enjoys similar windowing operations but makes use of four SA blocks grouped into two Cross Attention (CA) modules. The functionality of the CA can be described as:
\[ CA_r = \text{SA}(Q_D, K_D, V_D), \]  \tag{4}
\[ CA_l = \text{SA}(Q_D, K_E, V_E). \]  \tag{5}

The right branch of the CA acts on tokens generated by the previous VT-Dec-Blk according to Eq. (4). We emphasize on the flow of information from the decoder by the subscript \( D \) therein. The left branch of the CA, however, uses the queries generated by the decoder along with the keys and values obtained from the VT-Enc-Blk at the same level in the computation graph. The idea here is to use the basis spanned by the encoder (which is identified by values) along with its keys to benefit from spatial information harvested by the encoder. CA blocks, as mentioned above, also use the regular and shifted windowing to inject more inductive bias into the model. Note that the values and keys from the SA with the same windowing operation should be combined, hence the criss-cross connection form in Fig. 2.

**Remark 1.** One may ask why values and keys are considered from the encoder. We indeed studied other possibilities such as employing queries and keys from the encoder while generating values by the decoder. Empirically, the form described in Eq. (5) is observed to deliver better and more robust outcomes and hence our choice in VT-UNet.

**Remark 2.** We empirically observed that employing keys and values from the encoder in CA yields faster convergence of VT-UNet. This, we conjecture, is due to having extra connections from the decoder to encoder during the back-propagation which might facilitate gradient flow.

**Fusion Module.** As illustrated in Fig. 2b, tokens generated from the CA module and MSA module are combined together and fed to the next VT-Dec-Blk. \( \hat{z}_l \) is calculated using by a linear function as:

\[ \hat{z}_l = \alpha \hat{z}_l^c + (1 - \alpha) \hat{z}_l^s + F(\hat{z}_l^s), \]  \tag{6}

where \( F(\cdot) \) denotes Fourier Feature Positional Encoding (FPE) and \( \alpha \) controls the contribution from each CA and MSA module. In our experiments and for the sake of simplicity, we opt for \( \alpha = 0.5 \).

**Breaking the Symmetry.** Aiming for simplicity, in fusing tokens generated by the CA and MSA, we use a linear combination with \( \alpha = 0.5 \). This results in a symmetry, meaning that swapping \( \hat{z}_l^c \) and \( \hat{z}_l^s \) does not change the output. To break this symmetry and also better encapsulate object-aware representations that are critical for anatomical pixel-wise segmentation, we supplement the tokens generated from MSA by the 3D FPE. The 3D FPE employs sine and cosine functions with different frequencies Vaswani et al. (2017) to yield a unique encoding scheme for each token. The main idea is to use a sine/cosine function with a high frequency and modulate it across the dimensionality of the tokens while changing the frequency according to the location of the token within the 3D volume. We relegate the details to the supplementary material due to the lack of space.

**Classifier Layer.** After the final 3D patch expanding layer in the decoder, we introduce a classifier layer which includes a 3D convolutional layer to map deep \( C \) dimensional features to \( K \) segmentation classes. Therefore, the resulting prediction is a volume of \( D \times H \times W \times K \).
2.3 Training VT-UNet

The VT-UNet design provides a unified network architecture that could benefit various problems with volumetric data. But in this work we focus on medical image segmentation and the training process is more deeply shared across the loss function and architecture variants that we used.

**Loss Function.** Let $\mathcal{X} = \{(X_1, Y_1), \cdots, (X_n, Y_n)\}$ denote the labeled data from $n$ patients, where each pair $(X_i, Y_i)$ has an image $X_i \in \mathbb{R}^{D \times H \times W \times C}$ and its associated ground-truth mask $Y_i \in \mathbb{R}^{D \times H \times W \times K}$. To train VT-UNet, we jointly minimize the Dice Loss (DL) and Cross Entropy (CE) loss. The two losses are modified and computed in a voxel-wise manner. The DL is defined as:

$$\mathcal{L}_{\text{dl}}(\theta; \mathcal{X}) = -\mathbb{E}_{(X, Y) \sim \mathcal{X}} \left[ \frac{2(Y \cdot \mathcal{H}(X))}{|Y|_1 + |\mathcal{H}(X)|_1} \right],$$

where $\mathcal{H}(\cdot)$ and $\theta$ denote the transformer model and the model parameters, respectively. The CE loss is defined as:

$$\mathcal{L}_{\text{ce}}(\theta; \mathcal{X}) = \mathbb{E}_{(X, Y) \sim \mathcal{X}} \left[ -Y \log \mathcal{H}(X) \right],$$

Therefore, the total segmentation loss is:

$$\mathcal{L}_{\text{seg}}(\theta; \mathcal{X}) = \mathcal{L}_{\text{dl}}(\theta; \mathcal{X}) + \mathcal{L}_{\text{ce}}(\theta; \mathcal{X}).$$

**Architecture Variants.** We introduce variants of the VT-UNet, by changing the number of embedded dimensions used for model training. Our variants are:

- The tiny model: **VT-UNet-T** with $C = 48$
- The small model: **VT-UNet-S** with $C = 72$
- The base model: **VT-UNet-B** with $C = 96$

where $C$ is the embedded dimensions in the first stage. For all variants, the layer numbers for encoder, bottleneck and decoder are $\{2, 2, 2\}$, $\{1\}$, $\{2, 2, 2\}$, respectively. Initial patch partition uses window size of $P = 4$ and $M = 4$. The model size and FLOPs comparison is shown in Tab. 1.

3. Related Work

**CNN Based Methods.** Most of the recent deep CNN based approaches for image segmentation from 2D images have been built upon the seminal U-Net Ronneberger et al. (2015) architecture, which is a fully-convolutional encoder-decoder structure. Examples include U-Net++ Li et al. (2018), H-Dense-UNet Li et al. (2018), Res-UNet Xiao et al. (2018), Attention-UNet Oktay et al. (2018) and U-Net:3+ Huang et al. (2020). These methods have shown their success in segmentation and other tasks such as reconstruction Hyun et al. (2018); Han et al. (2019). The U-Net architecture was first extended to process 3D images by incorporating 3D convolutions in 3D UNet Çiçek et al. (2016). Later, V-Net Milletari et al. (2016) was proposed which develops a volumetric convolutional neural network with different stages with multiple resolutions. V-Net Milletari et al. (2016) adapts a learnable
residual function in each stage to improve the computational complexity of the network. Both 3D UNet Çiçek et al. (2016) and V-Net Milletari et al. (2016) can directly process the 3D medical images, and have demonstrated tremendous success in the field of medical AI due to their strong feature representation learning capability across multiple views over multiple slices.

**Transformer Based Methods.** Vision Transformers have shown promising empirical results for different computer vision tasks Touvron et al. (2021); Strudel et al. (2021); Carion et al. (2020); Zhu et al. (2020); Girdhar et al. (2019); Arnab et al. (2021); Liu et al. (2021b), with promising characteristics e.g. compared with the CNNs, they are less biased towards texture Naseer et al. (2021), show better generalization and robustness Shao et al. (2021); Naseer et al. (2021); Paul and Chen (2021); Mao et al. (2021); Bhojanapalli et al. (2021). Transformers have also been recently investigated for image segmentation Zheng et al. (2021); Chen et al. (2021b); Zhang et al. (2021); Valanarasu et al. (2021); Cao et al. (2021). The recent work of Kim et al. Kim et al. (2020), known as a Volumetric Transformer Network (VTN), is a neural model proposed to predict channel-wise warping fields and automatically localize discriminative objects using attention mechanism in vision regime which is an interesting study. TransUNet Chen et al. (2021b) is the first Transformer based approach for medical image segmentation. It adapts a UNet structure, and replaces the bottleneck layer with ViT Dosovitskiy et al. (2020) where patch embedding is applied on a feature map generated from CNN encoder (where input is a 2D slice of 3D volume). Different variants of TransUNet have been proposed Chang et al. (2021); Chen et al. (2021a); Zhang et al. (2021), where Convolutional blocks are used as the primary feature extractor with Transformer blocks. Unlike these hybrid approaches (using both convolutions and self-attention), Cao et al. Cao et al. (2021) proposed Swin-UNet, a purely transformer based network for medical image segmentation. It inherits swin-transformer blocks Liu et al. (2021a) and shows significant improvement over TransUNet Chen et al. (2021b). DS-TransUNet Lin et al. (2021) extends Swin-UNet Cao et al. (2021), by using dual Swin Transformer blocks in the decoder Lin et al. (2021).

There are some recent efforts to extend Transformer based models for 3D medical image segmentation Zhou et al. (2021); Hatamizadeh et al. (2021); Wang et al. (2021); Xie et al. (2021). The 3D version of TransUnet Chen et al. (2021b), called TransBTS Wang et al. (2021) has a CNN encoder-decoder design and a Transformer as the bottleneck layer. Zhou et al. Zhou et al. (2021) proposed nnFormer with 3D Swin Transformer blocks as encoder and decoder with interleaved stem of convolutions. A model which employs a transformer as the encoder and directly connects intermediate encoder outputs to the decoder via skip connections is proposed in Hatamizadeh et al. (2021). Yutong et al. Xie et al. (2021) introduced a hybrid model of CNN Transformer with a Deformable Transformer(DeTrans) employing the deformable SA mechanism. While these Transformer based approaches for 3D medical image segmentation have shown their promises, by achieving better performances compared with their CNN counterparts, they process the volumetric data by breaking it down into 2D slices, thus failing to encapsulate the full voxel information. Our proposed model, on the other hand, processes the volumetric data in its entirety, thus fully encoding the interactions between slices. Our proposed model is built on Transformers and introduces lateral connections to perform CA alongwith SA in the encoder-decoder design. These
design elements contributed to achieving better segmentation performance, along with enhanced robustness and generalization of features learned by our model.

4. Experiments

Table 1: Segmentation Results on BraTS 2021 Data. ↑ means higher is better.

| Method                  | #param. | FLOPS | Dice Score ↑ | Hausdorff Distance ↓ |
|-------------------------|---------|-------|---------------|-----------------------|
| 3D U-Net Čiçek et al. (2016) | 11.9 M  | 557.9 G | 83.39 - 86.28 - 89.59 - 86.32 | 6.15 - 6.18 - 11.49 - 7.94 |
| V-Ne Milletari et al. (2016) | 69.3 M  | 765.9 G | 81.04 - 84.71 - 90.32 - 85.36 | 7.53 - 7.48 - 17.20 - 10.73 |
| TransBTS Wang et al. (2021) | 33 M    | 333 G  | 80.35 - 85.35 - 89.25 - 84.99 | 7.83 - 8.21 - 15.12 - 10.41 |
| UNETR Hatamizadeh et al. (2021) | 102.5 M  | 193.5 G | 79.78 - 83.66 - 90.10 - 84.51 | 9.72 - 10.01 - 15.99 - 11.90 |
| nnFormer Zhou et al. (2021) | 39.7 M  | 110.7 G | 82.83 - 86.48 - 90.37 - 86.56 | 8.00 - 7.89 - 11.66 - 9.18 |
| VT-UNet-T | 5.4 M  | 52 G  | 83.04 - 86.68 - 90.48 - 86.82 | 9.46 - 9.23 - 13.34 - 10.68 |
| VT-UNet-S | 11.8 M  | 100.8 G | 83.14 - 86.86 - 91.02 - 87.00 | 8.25 - 8.03 - 11.46 - 9.25 |
| VT-UNet-B | 20.8 M  | 165.0 G | 85.59 - 87.41 - 91.20 - 88.07 | 6.23 - 6.29 - 10.03 - 7.52 |

Figure 4: Qualitative Results on BraTS 2021 Data. **Row-1**: yellow, red and white represent the peritumoral edema (ED), Enhancing Tumor (ET) and non enhancing tumor/necrotic tumor (NET/NCR), respectively. **Row-2**: segmentation boundaries. **Row-3**: volumetric tumor predicted by each method.

4.1 Experimental Settings

**BraTS Dataset.** We use the 1251 MRI scans of shape 240 × 240 × 155 from Multi-modal Brain Tumor Segmentation Challenge (BraTS) 2021 Baid et al. (2021); Menze et al. (2014); Bakas et al. (2017c,a,b). Following Ranjbarzadeh et al. (2021), we divide 1251 scans into 834, 208, 209 for training, validation and testing, respectively. We choose BraTS as our primary dataset because it reflects real world scenarios and exhibits diversity among MRI scans as they are acquired at various institutions with different equipment and protocols. BraTS dataset contains four distinct tumor sub-regions: (1) The Enhancing Tumor (ET),
Table 2: Ablation Study on individual components.

| Method             | Dice Score ↑ |
|--------------------|--------------|
|                   | ET  | TC  | WT  | AVG.|
| VT-UNet-T w/o CA & FPE |   80.81 | 82.35 | 88.59 | 83.92 |
| VT-UNet-S w/o CA & FPE |   82.62 | 85.75 | 90.49 | 86.29 |
| VT-UNet-B w/o CA & FPE |   83.50 | 87.34 | 90.64 | 87.16 |
| VT-UNet-T w/o FPE   |   82.51 | 86.70 | 89.95 | 86.39 |
| VT-UNet-S w/o FPE   |   82.76 | 86.99 | 90.69 | 86.81 |
| VT-UNet-B w/o FPE   |   84.14 | 87.37 | 90.92 | 87.47 |

(2) Non Enhancing Tumor (NET), (3) Necrotic Tumor (NCR) which along with NET, (4) Peritumoral Edema (ED). These almost homogeneous sub-regions can be clustered together to compose three semantically meaningful tumor classes: (1) Enhancing Tumor (ET), (2) the Tumor Core (TC) region (addition of ET, NET and NCR), and (3) the Whole Tumor (WT) region (addition of ED to TC).

Other Datasets. We also evaluate our method on Pancreas and Liver datasets from Medical Segmentation Decathlon (MSD) Antonelli et al. (2021); Simpson et al. (2019) for the task of tumor segmentation. Pancreas dataset has 281 CT volumes, divided into 187, 47, 47 for training, validation and testing, respectively. Liver dataset has 131 CT volumes, divided into 87, 22, 22 CT for training, validation and testing, respectively. Further, to evaluate the transferability of our pretrained features, we use MSD BraTS dataset having 484 MRI volumes, divided into 322, 81, 81 for training, validation and testing, respectively.

Implementation Details. We use PyTorch Paszke et al. (2017), with a single Nvidia RTX 3090 GPU. The weights of Swin-T Liu et al. (2021a) pre-trained on ImageNet-1K are used to initialize the model. For training, we employ Adam optimizer with a learning rate of $10^{-4}$ for 300 epochs using a cosine decay learning rate scheduler and a batch size of 1. To standardize all volumes, we perform min-max scaling followed by clipping intensity values, and cropping the volumes to a fixed size of $128 \times 128 \times 128$ by removing unnecessary background.

4.2 Experimental Results

Brain Tumor Segmentation. Tab. 1 compares our method with recent transformer based approaches (TransBTS Wang et al. (2021), UNETR Hatamizadeh et al. (2021) & nnFormer Zhou et al. (2021)) and two pioneering CNN based methods (3D UNet Çiçek et al. (2016) & V-Net Milletari et al. (2016)) for brain tumor segmentation on BraTS dataset. We report these results on test set of BraTS with 209 MRI scans. At inference, we use sliding window technique to generate multi-class predictions Ginsti et al. (2013); Gouk and Blake (2014). We select the model with best results on validation set for our evaluation. We use Dice Sørensen coefficient (Dice Score) and Hausdorff Distance (HD) as evaluation metrics, and separately compute them for three classes (i.e., ET, TC and WT), following similar evaluation strategy as in Jiang et al. (2019). Our quantitative results in Tab. 1 suggest that our approach consistently achieves best overall performance with Dice Score gains of 1.7%. Our method improves over nnFormer Zhou et al. (2021), the closely
competing hybrid approach, for Dice Score from 86.56% to 88.07% and HD from 9.18 to 7.52. Fig. 4 shows qualitative segmentation results on the test samples of previously unseen patient data. We can observe that our model can accurately segment the structure and delineates boundaries of tumor. We believe that, capturing long-range dependencies across adjacent slices plays a vital role in our model’s performance.

**Pancreas and Liver Tumor Segmentation.** To study generalization of our model, we further evaluate it on two other datasets (i.e., MSD Pancreas & MSD Liver). We compare our model’s performance with 3D UNet Çiçek et al. (2016) and nnFormer Zhou et al. (2021) in Tab. 6. We note that due to smaller size of training data, the overall performance for all methods is low on these datasets. Nevertheless, our proposed VT-UNet still performs better than the compared methods, when segmenting tumor, which is the most crucial class.

| Artfact  | VT-UNet-B | 3D UNet Çiçek et al. (2016) |
|----------|-----------|-----------------------------|
|          | ET        | TC  | WT  | AVG. | ET  | TC  | WT  | AVG. |
| Motion   | 75.13     | 78.96 | 85.78 | **79.96** | 69.86 | 76.74 | 77.44 | 74.68 |
| Ghost    | 74.50     | 79.25 | 84.88 | **79.54** | 72.57 | 78.88 | 78.19 | 76.55 |
| Spike    | 78.76     | 81.49 | 86.56 | **82.27** | 72.69 | 76.26 | 83.26 | 77.40 |
| Clean    | 85.59     | 87.41 | 91.20 | 88.07 | 83.39 | 86.28 | 89.59 | 86.42 |

Table 3: Dice Scores (higher the better) for different artefacts.

| Method                | Dice Score ↑ | Hausdorff Distance ↓ |
|-----------------------|--------------|-----------------------|
|                       | ET  | TC  | WT  | AVG. | ET  | TC  | WT  | AVG. |
| 3D UNet Çiçek et al. (2016) | 27.46 | 29.35 | 40.14 | 32.32 | 71.49 | 66.07 | 67.26 | 67.92 |
| VT-UNet-B             | **50.69** | **53.60** | **59.68** | **54.65** | **61.98** | **60.83** | **53.11** | **58.64** |

Table 4: Ablation Study on features transfer.

| Method | Pre-trained | ET  | TC  | WT  | AVG. |
|--------|-------------|-----|-----|-----|------|
| VT-UNet-B | ✅          | 85.59 | 87.41 | 91.20 | 88.07 |
| VT-UNet-B | ✗          | 83.26 | 86.82 | 91.06 | 87.05 |

Table 5: Dice Scores for pre-trained weights.

**4.3 Ablation Study & Analysis**

Here, we progressively integrate different components into the model to investigate their individual contribution towards overall performance. Our empirical results in Tab. 2 reveal the importance of introducing Parallel CA and SA together with FPE in VT-Dec-Blks along with convex combination. We can notice that all of these components contribute towards model’s performance. We also observe improved segmentation performance with an increased embedding dimensions of the feature space (i.e. variants of our model). Further, from Tab. 5, we can see that using pre-trained weights is helpful and boosts models performance.

**Generalization Analysis.** Here, we test the generalization capability of the features learned with our model, once they are transferred to other dataset. For this, we evaluate
Method & Liver & Pancreas & Liver & Tumor & AVG. & Pancreas & Tumor & AVG. \\
\hline
3D UNet & Çiçek et al. (2016) & 92.67 & 34.92 & 63.80 & 73.22 & 19.52 & 46.37 \\
nnFormer & Zhou et al. (2021) & 89.43 & 31.84 & 60.63 & 66.42 & 16.27 & 41.34 \\
VT-UNet-B & & 92.84 & 35.69 & 64.26 & 70.37 & 24.40 & 47.39 \\
\hline
| Table 6: Dice Scores for Segmentation results on other datasets. |

our model on MSD BRATS dataset, using features extracted from the model pre-trained on BraTS 2021 dataset. We only train the classifier on MSD BRATS dataset. For comparison, we consider 3D UNet Çiçek et al. (2016), and follow an identical evaluation setup. Our results in Tab. 4 suggest that features learned by our model generalize significantly better across dataset, compared with a CNN-based approach.

**Robustness Analysis.** Factors such as patient’s movement and acquisition conditions can introduce noise to MRI scans. Here, we investigate the robustness of our proposed approach, by synthetically introducing artefacts to MR images at inference time. These include (1) Motion artefacts, which can occur due to patient movements Shaw et al. (2018). (2) MRI ghosting artefacts, also known as phase-encoded motion artifacts e.g. cardiac and respiratory motions Axel et al. (1986). (3) MRI spike artefacts (Herringbone artefact), due to the spikes in k-space. Jin et al. (2017). Tab. 3 compares the robustness of our method with 3D UNet Çiçek et al. (2016). The results suggest that our approach performs more reliably in the presence of these nuisances. Our findings are consistent with existing works on RGB images, where Transformer based models have shown better robustness against occlusions Naseer et al. (2021), natural and adversarial perturbations Shao et al. (2021); Naseer et al. (2021); Paul and Chen (2021); Mao et al. (2021); Bhojanapalli et al. (2021), owing to their highly dynamic and flexible receptive field.

5. Conclusion & Limitations

This paper presents a volumetric transformer network for medical image segmentation, that is computationally efficient to handle large-sized 3D volumes, and learns representations that transfer better across datasets and are robust against artefacts. Our results show that the proposed model achieves consistent improvements over existing state-of-the-art methods in volumetric segmentation of tumors. We believe our work can assist better clinical diagnosis and treatment planning. We note that the public domain datasets considered in our work may lack diversity across ethnicity and race. We further note that the current deep learning methods, including ours, show lower results on datasets with fewer training samples. We leave volumetric segmentation with limited samples to future research.
Appendix A. Supplementary

This appendix provides additional details that were omitted from the main paper, due to the space restrictions.

A.1 Method

VT-Dec-Blk Functionality. In the VT-Dec-Blk, functionality continues as follow:

\[ \hat{z}_s^l = \text{VT-W-MSA} \left( \text{LN} \left( z_s^{l-1} \right) \right) + z_s^{l-1}, \]
\[ z_s^l = \text{MLP} \left( \text{LN} \left( \hat{z}_s^l \right) \right) + \hat{z}_s^l, \]
\[ \hat{z}_{s+1}^l = \text{VT-SW-MSA} \left( \text{LN} \left( z_s^l \right) \right) + z_s^l, \]
\[ z_{s+1}^l = \text{MLP} \left( \text{LN} \left( \hat{z}_{s+1}^l \right) \right) + \hat{z}_{s+1}^l, \]  \hspace{1cm} (10)

\[ \hat{z}_c^l = \text{VT-W-MSA} \left( \text{LN} \left( z_{c+1}^{l-1} \right) \right) + z_{c+1}^{l-1}, \]
\[ z_c^l = \text{MLP} \left( \text{LN} \left( \hat{z}_c^l \right) \right) + \hat{z}_c^l, \]
\[ \hat{z}_{c+1}^l = \text{VT-SW-MSA} \left( \text{LN} \left( z_c^l \right) \right) + z_c^l, \]
\[ z_{c+1}^l = \text{MLP} \left( \text{LN} \left( \hat{z}_{c+1}^l \right) \right) + \hat{z}_{c+1}^l, \]  \hspace{1cm} (11)

where \( \hat{z}_s^l \) and \( z_s^l \) denote the output features of the VT-W-MSA module and the MLP module for block \( l \), respectively.

![Fusion Module Diagram](image)

Figure 5: Fusion Module. In the VT-Dec-Blk, outputs of CA module (\( z_c^l \)) and MSA module (\( z_s^l \)) are fed to the Fusion module. \( z_c^l \) and \( z_s^l \) are then combined together using a convex combination approach. As a regularization and positional encoding method additional 3D Fourier Feature Positional Encoding (FPE) is introduced during combination.

Convergence We empirically observed that employing keys and values from the encoder in CA yields faster convergence of VT-UNet. This, we conjecture, is due to having extra
connections from the decoder to encoder during the back-propagation which might facilitate gradient flow. Fig. 6 shows the convergence of Dice Scores during validation phase for each of tumor class throughout first 40 epochs. For comparison we provide convergence charts for Proposed model and 3D UNet Çiçek et al. (2016).

**Figure 6: Convergence during Validation Phase.**

### Breaking Symmetry Cont.

The linear patch-projection flattens the features, thereby failing to fully encapsulate object-aware representations (e.g., spatial information) that are critical for anatomical pixel-wise segmentation. As shown in Fig. 5, combining two sets of tokens may result in loss of fluency. Therefore, in order to preserve features among continuous slices, we supplement the 3D Fourier Feature Positional Encoding (FPE) for the tokens generated from MSA module by adapting sine and cosine functions with different frequencies Vaswani et al. (2017), which provides unique encoding for each input token.

Following work by Wang *et al.* Wang and Liu (2021), we used the extended version of 2D positional encoding for 3D. Therefore, we call this as a 3D FPE or in other words a 3D positional encoding with a sinusoidal input mapping for VT-Dec-Blks. After applying 3D FPE to tokens, we pass it through a LN and MLP layer. Our empirical evaluations confirm that adding a Fourier feature positional bias can improve the poor conditioning of
the feature representation. We compute 3D FPE for each token as:

\[ \text{PE}( (x, y, z), 2i) = \sin(x \cdot \omega), \]
\[ \text{PE}( (x, y, z), 2i + 1) = \cos(x \cdot \omega), \]
\[ \text{PE}( (x, y, z), 2j + \frac{C}{3}) = \sin(y \cdot \omega), \]
\[ \text{PE}( (x, y, z), 2j + 1 + \frac{C}{3}) = \cos(y \cdot \omega), \]
\[ \text{PE}( (x, y, z), 2k + \frac{2C}{3}) = \sin(z \cdot \omega), \]
\[ \text{PE}( (x, y, z), 2k + 1 + \frac{2C}{3}) = \cos(z \cdot \omega), \]  
(12)

where \((x, y, z)\) is the position or a point in a 3D space and \(C\) is the dimensionality of the tokens. Here, \(\text{dim} = \{i, j, k\}\), where \(i,j,k\) is an integer in \([0, C/6)\) and \(\omega = 1/10000 \times \frac{\text{dim}}{C}\). For each input token, we get a positional encoding as the 3D positional encoding. We use a timescale ranging from 1 to 10000. The number of different timescales is equal to \(C/6\), corresponding to different frequencies. For each of these frequencies, we produce a sine/cosine signal as above on the horizontal/vertical direction. All these generated signals are then concatenated to \(C\) dimensions. The positional encoding and the token generated from Convex Combination are added together element wise and then forward pass to the next VT-Dec-Blk.

Figure 7: Visual Analysis of BraTS 2021 Training Data.

Figure 8: Qualitative Results on BraTS 2021 Data. Yellow, red and white represent the peritumoral edema (ED), Enhancing Tumor (ET) and non enhancing tumor/necrotic tumor (NET/NCR), respectively.
Figure 9: **Qualitative Results on MSD Liver Data.** Red and white represent the Liver organ and tumor, respectively.

Figure 10: **Qualitative Results on MSD Pancreas Data.** Red and white represent the Pancreas organ and tumor, respectively.

| Method | Dice Score ↑ | Hausdorff Distance ↓ |
|--------|---------------|----------------------|
|        | ET | TC | WT | AVG. | ET | TC | WT | AVG. |
| w/o CA & FPE | VT-UNet-T | 80.81 | 82.35 | 88.39 | 83.92 | 13.24 | 13.53 | 25.72 | 17.50 |
| VT-UNet-S | 82.62 | 86.75 | 90.49 | 86.29 | 11.26 | 10.40 | 13.04 | 11.57 |
| VT-UNet-B | 83.50 | 87.34 | 90.64 | 87.16 | 8.74 | 7.41 | 12.10 | 9.42 |
| w/o FPE | VT-UNet-T | 82.51 | 86.70 | 89.95 | 86.39 | 10.38 | 10.31 | 13.44 | 11.38 |
| VT-UNet-S | 82.76 | 86.99 | 90.69 | 86.81 | 9.64 | 8.81 | 13.39 | 10.61 |
| VT-UNet-B | 84.14 | 87.37 | 90.92 | 87.47 | 7.84 | 7.29 | 13.16 | 9.43 |
| with FPE | VT-UNet-T | 83.04 | 86.58 | 90.48 | 86.82 | 9.46 | 9.23 | 13.34 | 10.68 |
| VT-UNet-S | 83.14 | 86.86 | 91.02 | 87.00 | 8.25 | 8.03 | 11.46 | 9.25 |
| VT-UNet-B | 85.59 | 87.41 | 91.20 | 88.07 | 6.23 | 6.29 | 10.03 | 7.52 |

Table 7: Ablation Study on individual components.

### A.2 Experiments

**More Details on BraTS Dataset.** We use the MRI scans of 1251 MR volumes of shape 240 × 240 × 155 from Multi-modal Brain tumour Segmentation Challenge (BraTS) 2021 Baid et al. (2021); Menze et al. (2014); Bakas et al. (2017c,a,b). Following previous works Ranjbarzadeh et al. (2021), we divide 1251 scans into 834, 208, 209 for training, validation and testing, respectively. We choose, BraTS 2021 dataset as our primary dataset because it has multi-institutional and multi-parametric MRI scans, and is considered one of the largest and most common bench-marking venue in medical AI domain. The dataset reflects real world scenarios, and exhibits diversity among MRI volumes since scans are acquired from various institutions using different equipment and different protocols.

These MRI sequences are conventionally used for glioma detection: T1 weighted sequence (T1), T1-weighted contrast enhanced sequence using gadolinium contrast agents (T1Gd) (T1CE), T2 weighted sequence (T2), and Fluid attenuated inversion recovery
Motion Artefact
Spike Artefact
Ghost Artefact
Input Image Input Image
Ground Truth VT-UNet-B w/o artefacts
VT-UNet-B with artefacts

Figure 11: Qualitative Results Generated For Artefacted Data. Yellow, red and white represent the peritumoral edema (ED), Enhancing Tumor (ET) and non enhancing tumor/necrotic tumor (NET/NCR), respectively.

Input Flair Ground Truth VT-UNet-B 3D-UNet

Figure 12: Qualitative Results on features transfer. Yellow, red and white represent the peritumoral edema (ED), Enhancing Tumor (ET) and non enhancing tumor/necrotic tumor (NET/NCR), respectively.

(FLAIR) sequence. Using these sequences, four distinct tumor sub-regions can be identified from MRI as: (1) The Enhancing Tumor (ET) which corresponds to area of relative hyper-intensity in the T1CE with respect to the T1 sequence, (2) Non Enhancing Tumor (NET), (3) Necrotic Tumor (NCR) which along with NET is hypo-intense in T1-Gd when compared to T1, (4) Peritumoral Edema (ED) which is hyper-intense in FLAIR sequence. These almost homogeneous sub-regions can be clustered together to compose three semantically meaningful tumor classes: (1) Enhancing Tumor (ET), (2) the Tumor Core (TC) region (addition of ET, NET and NCR), and (3) the Whole Tumor (WT) region (addition of ED to TC). MRI sequences and ground truth map with three classes are shown in Fig. 7.

More Qualitative Comparisons. Here we present more qualitative results for the comparison. Fig. 8, Fig. 9 and Fig. 10 show segmentation results generated by each method including ours for BraTS 2021 dataset, MSD Liver dataset and MSD Pancreas dataset, re-
Figure 13: The box and whisker plot of the distribution for Dice Scores of 209 Unseen Patient Data.

| Tumor Class | Mean | Median | Standard Deviation |
|-------------|------|--------|--------------------|
| ET          | 85.59 | 90.41  | 8.82               |
| TC          | 87.41 | 94.14  | 6.55               |
| WT          | 91.20 | 93.75  | 4.75               |

Table 8: Statistical Insights From Box-Plot.

spectively. From the resultant segmentation maps, it can be seen that the model has shown a consistent performance over multiple datasets with multiple modalities (i.e., MRI, CT).

**More Quantitative Comparisons.** In this section, we provide a comparison on model performance with and without important design elements for all three variants (i.e., Tiny, Small, Base). Evaluation is conducted based on Dice Score and Hausdorff distance as shown in Tab. 7.

**Generalization Analysis Cont.** Here, we test the generalization capability of the features learned with our model, once they are transferred to other dataset. In the main paper showed the quantitative analysis for generalization ability of our proposed model and 3D UNet Çiçek et al. (2016) for MSD BraTS dataset. Here, we show some visualizations of generated resultant segmentation masks in Fig. 12.

**Robustness Analysis Cont.** In the main paper we investigated the robustness of our proposed approach, by synthetically introducing artefacts to MR images at inference time and shown quantitative analysis. In this section we provide more visualizations on artefacts we added during inference and the generated predictions to show the robustness of the proposed model in Fig. 11. It can be seen that, the predictions generated for artefacted
data are visually closer to the predictions generated without any artefacts. The model’s robustness can be further enhanced by adding artefacts during model training which we will consider in our future works.

**Statistical Insights.** Here we provide some statistical insights of the experimental results for unseen patient data evaluated during testing phase for VT-UNet-B model. Fig. 13 and Tab. 8 shows the distribution of dice scores for unseen cases. The box-plot shows the minimum, lower quartile, median, upper quartile and maximum for each tumor class. Outliers are shown away from lower quartile.

**References**

Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, Bram van Ginneken, et al. The medical segmentation decathlon. *arXiv preprint arXiv:2106.05735*, 2021.

Saeed Anwar, Salman Khan, and Nick Barnes. A deep journey into super-resolution: A survey. *ACM Computing Surveys (CSUR)*, 53(3):1–34, 2020.

Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. *arXiv preprint arXiv:2103.15691*, 2021.

L Axel, RM Summers, HY Kressel, and C Charles. Respiratory effects in two-dimensional fourier transform mr imaging. *Radiology*, 160(3):795–801, 1986.

Ujjwal Baid, Satyam Ghodasara, Michel Bilello, Suyash Mohan, Evan Calabrese, Errol Colak, Keyvan Farahani, Jayashree Kalpathy-Cramer, Felipe C Kitamura, Sarthak Pati, et al. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv preprint arXiv:2107.02314*, 2021.

Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin Kirby, John Freymann, Keyvan Farahani, and Christos Davatzikos. Segmentation labels and radiomic features for the pre-operative scans of the tcga-gbm collection. the cancer imaging archive. *Nat Sci Data*, 4:170117, 2017a.

Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin Kirby, John Freymann, Keyvan Farahani, and Christos Davatzikos. Segmentation labels and radiomic features for the pre-operative scans of the tcga-lgg collection. *The cancer imaging archive*, 286, 2017b.

Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S Kirby, John B Freymann, Keyvan Farahani, and Christos Davatzikos. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific data*, 4(1):1–13, 2017c.

Srinadh Bhojanapalli, Ayan Chakrabarti, Daniel Glasner, Daliang Li, Thomas Unterthiner, and Andreas Veit. Understanding robustness of transformers for image classification. *arXiv preprint arXiv:2103.14586*, 2021.
Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. *arXiv preprint arXiv:2105.05537*, 2021.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020.

Yao Chang, Hu Menghan, Zhai Guangtiao, and Zhang Xiao-Ping. Transclaw u-net: Claw u-net with transformers for medical image segmentation. *arXiv preprint arXiv:2107.05188*, 2021.

Bingzhi Chen, Yishu Liu, Zheng Zhang, Guangming Lu, and David Zhang. Transat-tunet: Multi-level attention-guided u-net with transformer for medical image segmentation. *arXiv preprint arXiv:2107.05274*, 2021a.

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

Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

Ozgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention*, pages 424–432. Springer, 2016.

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

Rohit Girdhar, Joao Carreira, Carl Doersch, and Andrew Zisserman. Video action transformer network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 244–253, 2019.

Alessandro Giusti, Dan C Cireşan, Jonathan Masci, Luca M Gambardella, and Jürgen Schmidhuber. Fast image scanning with deep max-pooling convolutional neural networks. In *2013 IEEE International Conference on Image Processing*, pages 4034–4038. IEEE, 2013.

Henry GR Gouk and Anthony M Blake. Fast sliding window classification with convolutional neural networks. In *Proceedings of the 29th International Conference on Image and Vision Computing New Zealand*, pages 114–118, 2014.

Yoseo Han, Leonard Sunwoo, and Jong Chul Ye. k-space deep learning for accelerated mri. *IEEE transactions on medical imaging*, 39(2):377–386, 2019.
Ali Hatamizadeh, Dong Yang, Holger Roth, and Daguang Xu. U ner: Transformers for 3d medical image segmentation. *arXiv preprint arXiv:2103.10504*, 2021.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

Han Hu, Zheng Zhang, Zhenda Xie, and Stephen Lin. Local relation networks for image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3464–3473, 2019.

Huimin Huang, Lanfen Lin, Ruofeng Tong, Hongjie Hu, Qiaowei Zhang, Yutaro Iwamoto, Xianhua Han, Yen-Wei Chen, and Jian Wu. Unet 3+: A full-scale connected unet for medical image segmentation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1055–1059. IEEE, 2020.

Chang Min Hyun, Hwa Pyung Kim, Sung Min Lee, Sungchul Lee, and Jin Keun Seo. Deep learning for undersampled mri reconstruction. *Physics in Medicine & Biology*, 63(13):135007, 2018.

Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, Olivier Hénaff, Matthew M. Botvinick, Andrew Zisserman, Oriol Vinyals, and João Carreira. Perceiver io: A general architecture for structured inputs & outputs, 2021a.

Andrew Jaegle, Felix Gimeno, Andrew Brock, Andrew Zisserman, Oriol Vinyals, and Joao Carreira. Perceiver: General perception with iterative attention. *arXiv preprint arXiv:2103.03206*, 2021b.

Zeyu Jiang, Changxing Ding, Minfeng Liu, and Dacheng Tao. Two-stage cascaded u-net: 1st place solution to brats challenge 2019 segmentation task. In *International MICCAI Brainlesion Workshop*, pages 231–241. Springer, 2019.

Kyong Hwan Jin, Ji-Yong Um, Dongwook Lee, Juyoung Lee, Sung-Hong Park, and Jong Chul Ye. Mri artifact correction using sparse+ low-rank decomposition of annihilating filter-based hankel matrix. *Magnetic resonance in medicine*, 78(1):327–340, 2017.

Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shah-baz Khan, and Mubarak Shah. Transformers in vision: A survey. *arXiv preprint arXiv:2101.01169*, 2021.

Seungryong Kim, Sabine Süssstrunk, and Mathieu Salzmann. Volumetric transformer networks. In *European Conference on Computer Vision*, pages 561–578. Springer, 2020.

Xiaomeng Li, Hao Chen, Xiaojuan Qi, Qi Dou, Chi-Wing Fu, and Pheng-Ann Heng. H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes. *IEEE transactions on medical imaging*, 37(12):2663–2674, 2018.
Ailiang Lin, Bingzhi Chen, Jiayu Xu, Zheng Zhang, and Guangming Lu. Ds-transunet: Dual swin transformer u-net for medical image segmentation. *arXiv preprint arXiv:2106.06716*, 2021.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021a.

Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. *arXiv preprint arXiv:2106.13230*, 2021b.

Xiaofeng Mao, Gege Qi, Yuefeng Chen, Xiaodan Li, Ranjie Duan, Shaokai Ye, Yuan He, and Hui Xue. Towards robust vision transformer. *arXiv preprint arXiv:2105.07926*, 2021.

Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993–2024, 2014.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. IEEE, 2016.

Muzammal Naseer, Kanchana Ranasinghe, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers. *arXiv preprint arXiv:2105.10497*, 2021.

Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

Sayak Paul and Pin-Yu Chen. Vision transformers are robust learners. *arXiv preprint arXiv:2010.07581*, 2021.

Ramin Ranjbarzadeh, Abbas Bagherian Kasgari, Saeid Jafarzadeh Ghoushchi, Shokofeh Anari, Maryam Naseri, and Malika Bendechache. Brain tumor segmentation based on deep learning and an attention mechanism using mri multi-modalities brain images. *Scientific Reports*, 11(1):1–17, 2021.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241. Springer, 2015.

Rulin Shao, Zhouxing Shi, Jinfeng Yi, Pin-Yu Chen, and Cho-Jui Hsieh. On the adversarial robustness of visual transformers. *arXiv preprint arXiv:2103.15670*, 2021.
Richard Shaw, Carole Sudre, Sebastien Ourselin, and M Jorge Cardoso. Mri k-space motion artefact augmentation: model robustness and task-specific uncertainty. In International Conference on Medical Imaging with Deep Learning–Full Paper Track, 2018.

Amber L Simpson, Michela Antonelli, Spyridon Bakas, Michel Bilello, Keyvan Farahani, Bram Van Ginneken, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjorn Menze, et al. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. arXiv preprint arXiv:1902.09063, 2019.

Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation, 2021.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In International Conference on Machine Learning, pages 10347–10357. PMLR, 2021.

Jeya Maria Jose Valanarasu, Poojan Oza, Ilker Hacihaliloglu, and Vishal M Patel. Medical transformer: Gated axial-attention for medical image segmentation. arXiv preprint arXiv:2102.10662, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Advances in neural information processing systems, pages 5998–6008, 2017.

Chengjia Wang, Tom MacGillivray, Gillian Macnaught, Guang Yang, and David Newby. A two-stage 3d unet framework for multi-class segmentation on full resolution image. arXiv preprint arXiv:1804.04341, 2018.

Wenxuan Wang, Chen Chen, Meng Ding, Hong Yu, Sen Zha, and Jiangyun Li. Transbts: Multimodal brain tumor segmentation using transformer. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 109–119. Springer, 2021.

Zelun Wang and Jyh-Charn Liu. Translating math formula images to latex sequences using deep neural networks with sequence-level training. International Journal on Document Analysis and Recognition (IJDAR), 24(1):63–75, 2021.

Dufan Wu, Kuang Gong, Kyungsang Kim, Xiang Li, and Quanzheng Li. Consensus neural network for medical imaging denoising with only noisy training samples. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 741–749. Springer, 2019.

Yingda Xia, Dong Yang, Zhiding Yu, Fengze Liu, Jinzheng Cai, Lequan Yu, Zhuotun Zhu, Daguang Xu, Alan Yuille, and Holger Roth. Uncertainty-aware multi-view co-training for semi-supervised medical image segmentation and domain adaptation. Medical Image Analysis, 65:101766, 2020.
Xiao Xiao, Shen Lian, Zhiming Luo, and Shaozi Li. Weighted res-unet for high-quality retina vessel segmentation. In 2018 9th international conference on information technology in medicine and education (ITME), pages 327–331. IEEE, 2018.

Yutong Xie, Jianpeng Zhang, Chunhua Shen, and Yong Xia. Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation. arXiv preprint arXiv:2103.03024, 2021.

Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. Cross-modal self-attention network for referring image segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10502–10511, 2019.

Li Yuan, Yinpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zihang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. arXiv preprint arXiv:2101.11986, 2021.

Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Learning enriched features for real image restoration and enhancement. In ECCV, 2020.

Yue Zhang, Shun Miao, Tommaso Mansi, and Rui Liao. Unsupervised x-ray image segmentation with task driven generative adversarial networks. Medical image analysis, 62:101664, 2020.

Yundong Zhang, Huiye Liu, and Qiang Hu. Transfuse: Fusing transformers and cnns for medical image segmentation. arXiv preprint arXiv:2102.08005, 2021.

Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6881–6890, 2021.

Hong-Yu Zhou, Jiansen Guo, Yinghao Zhang, Lequan Yu, Liangsheng Wang, and Yizhou Yu. nnformer: Interleaved transformer for volumetric segmentation. arXiv preprint arXiv:2109.03201, 2021.

Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, pages 3–11. Springer, 2018.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159, 2020.