DUViT-Net: Enhancing 2D DoubleU-Net with Vision Transformer for Medical Image Segmentation

Parthipan Ramakrishnan

January 6, 2025

Abstract

Medical image segmentation is pivotal in clinical applications such as diagnosis, treatment planning, and disease monitoring. Accurate segmentation enables precise measurement of anatomical and pathological regions, which is essential for effective patient care. This thesis presents DUViT-Net, a 2D DoubleU-Net architecture enhanced with dual Vision Transformer (ViT) blocks, evaluated on Task 01 (Brain Tumor Segmentation), Task 02 (Left Atrium Segmentation), and Task 04 (Hippocampus Segmentation) of the Medical Segmentation Decathlon (MSD) dataset. DUViT-Net combines convolutional neural networks (CNNs) for local feature extraction with transformers to capture global contextual information. The study includes comprehensive data preprocessing, model training with optimized hyperparameters, and evaluation using metrics like Dice coefficient, Jaccard index/Intersection over Union (IoU), Hausdorff Distance (HD95), F2 score, precision, and recall. Results demonstrate that DUViT-Net achieves up to a 7% improvement in the Dice coefficient over baseline models. Notably, DUViT-Net achieves a higher Jaccard index and recall than state-of-the-art models, indicating better overlap with ground truth and fewer false negatives. These findings highlight DUViT-Net's clinical potential to improve segmentation accuracy in resource-constrained settings.

1 Introduction

Background and Motivation: Medical image segmentation involves partitioning images into meaningful regions, such as anatomical structures or pathological areas. It is vital in clinical applications, including diagnostic imaging, treatment planning, surgical navigation, and disease monitoring. Accurate segmentation enables precise measurements of organ volumes, detection of abnormalities, and guidance for surgical procedures, which are essential for effective patient care. However, segmentation tasks face challenges such as anatomical variability, low

contrast between tissues, and noise, which complicate the segmentation process. The traditional methods often struggle to address these issues, necessitating advanced techniques to achieve improved precision in delineating boundaries. Recent advancements in deep learning, particularly convolutional neural networks (CNNs), have revolutionized image segmentation by enabling models to automatically learn hierarchical feature representations from raw data. Architectures like U-Net [27] and its variants are now widely used for their ability to capture multi-scale information through encoder-decoder structure [6]. These models have demonstrated superior performance in various medical imaging tasks, outperforming traditional segmentation methods. Recently, transformer architectures have shown remarkable success in natural language processing and have begun to make significant inroads in computer vision [7; 42] tasks. Vision Transformers (ViT) [9] adapt the transformer architecture for image data, effectively capturing long-range dependencies and global context. Unlike CNNs, which primarily focus on localized feature extraction, transformers excel at modelling relationships between distant parts of an image, potentially capable of enhancing CNN-based segmentation accuracy.

Problem Statement: While various models have been applied to the Medical Segmentation Decathlon (MSD) [3] dataset, as far as we know, there are no studies evaluating the integration of ViT blocks into the DoubleU-Net architecture specifically for 2D medical image segmentation. Previous works like TransUNet [5] and Swin UNETR [11] have explored the combination of CNNs and transformers, primarily in 3D contexts or with different architectural designs. This gap offers an opportunity to assess the effectiveness of combining CNNs with transformers in a 2D framework across multiple tasks and imaging modalities.

Importance of Evaluation: Evaluating the performance of our proposed model, a 2D DoubleU-Net with ViT blocks on the MSD dataset is crucial for understanding its generalizability and robustness across various tasks and imaging modalities. Improved segmentation accuracy can have significant clinical implications, such as enhancing the precision of tumour delineation in radiotherapy planning or improving the detection of small lesions indicative of early disease. In our study we will use several widely used metrics in our evaluation.

Objectives: This thesis investigates our proposed model, which integrates ViT blocks into a 2D DoubleU-Net architecture for medical image segmentation on the MSD dataset. The proposed model aims to leverage the local feature extraction capabilities of CNNs and the global context modelling of transformers to improve segmentation performance. By combining these strengths, the model seeks to achieve higher accuracy and better delineation of complex anatomical structures while maintaining computational efficiency.

Primary Objective: To implement and evaluate a 2D DoubleU-Net architecture enhanced with Vision Transformer blocks on tasks 01 (Brain Tumour Segmentation), 02 (Left Atrium Segmentation), and 04 (Hippocampus Segmentation) of the MSD dataset.

Secondary Objectives:

- Compare the performance of the proposed model with baseline models and state-of-the-art architectures using comprehensive evaluation metrics.
- Analyze the advantages and limitations of integrating ViT blocks into the DoubleU-Net architecture in the context of different medical imaging tasks.
- Explore the potential of incorporating other backbone architectures like ResNet50 and EfficientNet to enhance performance further.
- Provide insights into the applicability of the proposed model for various medical image segmentation applications, informing future research.

Contributions:

Novel Integration: Introducing DUViTNet, which integrates dual ViT blocks into a 2D DoubleUNet framework for the MSD dataset, providing a novel architecture applicable to multiple medical image segmentation tasks.

Comprehensive Evaluation: Utilizing a range of evaluation metrics beyond standard measures, this research offers an in-depth analysis of the proposed DUViT-Net model's performance, highlighting strengths and areas for improvement in segmentation accuracy and reliability.

2 Background and Related Work

Medical Image Segmentation Medical image segmentation refers to the process of outlining anatomical structures or pathological regions within medical images, such as MRI or CT scans. This segmentation is fundamental in aiding healthcare professionals with diagnosis, treatment planning, and monitoring disease progression. Accurate segmentation facilitates quantitative analysis, enabling precise measurements and assessments critical for effective patient care.

Traditional Methods: Traditional segmentation techniques include thresholding [26], region growing [1], active contours [21], and atlas-based methods [31]. Thresholding involves classifying pixels based on intensity values, while region growing expands regions from seed points based on predefined criteria. Active contours, or snakes, use energy minimization to detect object boundaries, and atlas-based methods rely on statistical shape models for segmentation.

Limitations of Traditional Methods: While traditional methods have been widely used, they often struggle with complex medical images. Challenges such as high variability in anatomical structures, low contrast between different tissues, noise, and the presence of artifacts can lead to inaccurate segmentation. These methods typically require extensive manual tuning and may not generalize well across different datasets or imaging modalities, limiting their applicability in diverse clinical settings.

2.1 CNNs in Medical Imaging

U-Net Architecture: Introduced by Ronneberger et al. [27], U-Net is a widely used CNN architecture for biomedical image segmentation. It features an encoder-decoder structure with skip connections, enabling precise localization and contextual understanding. The encoder progressively reduces spatial dimensions while increasing feature channels, capturing high-level features. The decoder restores spatial dimensions, combining them with corresponding encoder features through skip connections to maintain spatial accuracy.

*DoubleU-Net:*Proposed by Jha et al.[19], DoubleU-Net employs two U-Net architectures sequentially to refine segmentation outputs. The first U-Net performs coarse segmentation, then is refined by the second U-Net. This stacking enhances segmentation accuracy by correcting errors from the initial prediction and improving boundary delineation.

*ASPP Module:*The Atrous Spatial Pyramid Pooling (ASPP) module proposed by Chen et al. [6] is used to capture multi-scale information by applying parallel atrous convolutions, also known as dilated convolutions, with different dilation rates. This enhances the model's ability to handle objects of varying sizes without increasing computational complexity.

2.2 Vision Transformers

Transformer Architecture: Initially developed for sequence modelling in natural language processing by Vaswani et al. [36], transformers use self-attention mechanisms to capture global dependencies. Unlike CNNs, transformers can simultaneously model relationships between all parts of the input, making them effective for tasks requiring global context. Many transformer-based networks [34; 39; 40; 43] are proposed for medical image segmentation.

Vision Transformer (ViT): Dosovitskiy et al. [9] applied transformers to image classification by splitting images into fixed-size patches, embedding them, and treating the sequence of embeddings as input tokens to a transformer. ViTs achieved state-of-the-art results on image classification benchmarks, demonstrating the potential of transformers in computer vision.

ViT in Medical Imaging: In medical imaging, ViTs have been explored for segmentation tasks to capture long-range dependencies and improve the modelling of complex anatomical structures. Integrating ViT into existing architectures can enhance performance by leveraging local and global features [35]. However, our proposed DUViT-Net model goes a step further by integrating ViT blocks into the DoubleU-Net architecture, not just for global feature extraction but also to enhance the segmentation refinement process.

2.3 Hybrid CNN-Transformer Models

TransUNet: Chen et al. [4] introduced TransUNet, combining CNNs and transformers for medical image segmentation. The model uses a CNN encoder

to extract low-level features and a transformer to capture global context, demonstrating improved performance over purely CNN-based models.

Swin UNETR: Hatamizadeh et al. [11] proposed Swin UNETR, integrating Swin Transformers into a 3D U-Net architecture for volumetric medical image segmentation. Swin UNETR achieved state-of-the-art results on several datasets by effectively modelling both local and global features in 3D space.

2.4 State-of-the-Art Models

Advancements in medical image segmentation have been driven by the development of innovative architectures that enhance the accuracy and efficiency of segmentation tasks. This section reviews three prominent state-of-the-art models: DiNTS, nnU-Net, and MedVisionLlama. These models serve as benchmarks against which the proposed model DUViT-Net is evaluated.

DiNTS: Differentiable Neural Network Topology Search for 3D Medical Image Segmentation. He et al. [13] introduced DiNTS, which employs differentiable neural network topology search within a 3D segmentation framework. The model automates the discovery of optimal network architectures tailored for specific segmentation tasks, enhancing performance without manual intervention. *nnU-Net: Self-adapting Framework for U-Net-Based Medical Image Segmentation* proposed by Isensee et al. [16] stands out as a self-adapting framework that automates the configuration of U-Net architectures for various medical image segmentation tasks. The framework dynamically adjusts preprocessing steps, network architecture, and training procedures based on dataset characteristics. *MedVisionLlama: Leveraging Pre-Trained Large Language Model Layers to Enhance Medical Image Segmentation* proposed by [24] introduced MedVisionLlama, a novel approach that integrates pre-trained large language model (LLM) transformer blocks into ViT-based medical image segmentation models. The model leverages the pre-trained knowledge from extensive textual data to enhance feature representation by incorporating a frozen LLM transformer block into the encoder. However, these approaches either focus on 3D data [13] or require extensive computational resources, highlighting the need for a resource-efficient model like DUViT-Net, which utilizes a 2D architecture that is less resource-intensive than the 3D counterparts.

2.5 Summary of Related Work

Several studies have applied various deep learning models to the Medical Segmentation Decathlon dataset, including U-Net [27], ResNet-based [12] architectures, and attention [36] enhanced models. These studies have demonstrated improvements in segmentation accuracy and robustness, leveraging the strengths of different architectural components and training strategies[16]. Different architectures exhibit varying strengths and weaknesses across tasks. For instance, models with deeper encoders [12; 15] perform better in capturing complex features but require more computational resources. Attention mechanisms have

shown promise in focusing on relevant regions, enhancing segmentation precision [25; 28]. Despite extensive research to our knowledge, there are no previous studies evaluating the integration of ViT blocks into the DoubleU-Net architecture for 2D segmentation tasks. While MedVisionLlama (Kumar et al., 2024) [24] explores the integration of LLM transformer blocks into ViT-based models, there remains an opportunity to investigate the benefits of combining ViT blocks with DoubleU-Net in a 2D context. This gap highlights the need to assess the potential of DoubleU-Net with ViT blocks in handling the diverse and challenging tasks presented by the dataset.

3 Fundamentals

This section provides background information and the evaluation measures used to assess the performance of the methods.

3.1 Evaluation Metrics in Medical Segmentation

Accurate evaluation of segmentation models is crucial to understanding their performance and reliability in clinical settings. The following metrics are employed to assess the proposed model comprehensively:

- *Dice Coefficient:* The Dice coefficient [32] measures the overlap between the predicted segmentation and the ground truth. It is defined as twice the area of overlap divided by the total number of pixels in both the predicted and ground truth masks. A higher Dice score indicates better segmentation accuracy.

$$\text{Dice} = \frac{2|A \cap B|}{|A| + |B|} \tag{1}$$

  - A is the set of predicted positive pixels.
  - B is the set of ground truth positive pixels.

- *Jaccard Index/IoU:* The Jaccard Index, or Intersection over Union (IoU) [32], quantifies the similarity between the predicted and ground truth masks by dividing the intersection by the union of the two sets. It measures the shared area between the predictions and ground truth, with higher values indicating better performance.

$$\text{Jaccard} = \frac{|A \cap B|}{|A \cup B|} \tag{2}$$

  - A is the set of predicted positive pixels.
  - B is the set of ground truth positive pixels.

- *Precision and Recall:* Precision measures the proportion of correctly predicted positive pixels out of all predicted positive pixels, while recall measures the proportion of correctly predicted positive pixels out of all actual positive pixels. High precision reduces false positives, and high recall minimizes false negatives, both critical in medical image segmentation.

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

  Where TP, FP, and FN represent true positives, false positives, and false negatives, respectively.

- *F2 Score:* The F2 score [37; 38] is a weighted harmonic mean of precision and recall, giving more emphasis to recall. It is essential in medical contexts where missing a pathological region (false negatives) can have severe consequences. A higher F2 score [19] indicates better recall while maintaining reasonable precision.

$$\text{F2} = \frac{5 \cdot \text{Precision} \cdot \text{Recall}}{4 \cdot \text{Precision} + \text{Recall}} \tag{4}$$

- *HD95 (95th Percentile Hausdorff Distance):* HD95 [23] measures the distance between the boundaries of the predicted and ground truth masks, specifically the 95th percentile, to reduce the impact of outliers. It assesses the boundary agreement and robustness of the segmentation, with lower values indicating better performance.

$$\text{HD}(A, B) = \max \left\{ \sup_{a \in A} \inf_{b \in B} \|a - b\|, \sup_{b \in B} \inf_{a \in A} \|b - a\| \right\} \tag{5}$$

  where

  - A is the set of predicted positive pixels.
  - B is the set of ground truth positive pixels.

3.2 Loss Functions:

The combination of Dice loss and Binary Cross-Entropy (BCEWithLogitsLoss) was utilized to balance the optimization between overlapping regions and pixel-wise classification.

- *Dice Loss:* Dice Loss [19] is derived from the Dice coefficient, a metric traditionally used to gauge the similarity between two samples. In the context of segmentation, it quantifies the overlap between the predicted segmentation and the ground truth. The Dice coefficient D is defined as:

$$D = \frac{2 \times |A \cap B|}{|A| + |B|} \tag{6}$$

  where:

– A is the set of predicted positive pixels.
– B is the set of ground truth positive pixels.

To convert this metric into a loss function suitable for optimization, Dice Loss is formulated as:

$$DiceLoss = 1 - D = 1 - \frac{2 \times (A \cdot B)}{\parallel A \parallel_1 + \parallel B \parallel_1} \tag{7}$$

- *Binary Cross-Entropy with Logits Loss (BCEWithLogitsLoss)*: Binary Cross-Entropy (BCE) [19] loss measures the discrepancy between the predicted probabilities and the actual binary labels for each pixel. When combined with logits, it is referred to as BCEWithLogitsLoss. This loss function is defined as:

$$BCE = -\frac{1}{N} \sum_{i=1}^{N} [y_i \cdot \log(\widehat{y}_i) + (1 - y_i) \cdot \log(1 - \widehat{y}_i)] \tag{8}$$

where:

– N is the total number of pixels.
– $y_i \in \{1, ...N\}$ represents the ground truth label for i^{th} pixel.
– $\widehat{y}_i$ is the predicted probability (output of sigmoid activation) for the i^{th} pixel.

BCEWithLogitsLoss is effective for pixel-wise classification tasks, ensuring that each pixel is individually classified based on its ground truth label.

3.3 Training fundamentals

Optimization Algorithm: The Adam optimizer [22] was chosen for its adaptive learning rate capabilities. The model adapts learning rates for individual parameters based on the first and second moments of the gradients:

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{\hat{v}_t} + \epsilon} \hat{m}_t \tag{9}$$

where θ_t represents the parameters (weights and biases) of the model at time step t, $\hat{m}_t$ and $\hat{v}_t$ are the bias-corrected first and second moment estimates, and η is the learning rate.

Learning rate scheduler: ReduceLROnPlateau [2] is employed to reduce the learning rate by a factor of 0.1 if the validation loss does not improve for five consecutive epochs, facilitating convergence. These settings facilitate efficient and stable convergence during training.

Mixed Precision Training: Mixed precision training [10] leverages half-precision (float16) computations alongside single-precision (float32), reducing memory

consumption and accelerating training without compromising performance. In this approach, operations such as gradient calculations and activation updates are performed in float16, which allows for more efficient memory usage and faster computation. To ensure the stability of the training process, model weights and certain critical operations, such as weight updates, are maintained in float32 to prevent precision loss.

3.4 Data Augmentation

Data augmentation is crucial due to the limited availability of annotated medical images and the need for models to generalize across diverse clinical scenarios. As described below, these augmentations were applied to the training dataset to introduce variability and enhance the model's robustness.

3.4.1 Geometric Transformations

- *Horizontal and vertical flipping:* Applied to simulate variations in patient positioning. This is particularly relevant in modalities where anatomical structures are symmetric or can appear in different orientations (e.g., abdominal CT scans).
- *Rotation:* Random rotations within ± 15 degrees address slight misalignments during image acquisition [8].
- *Scaling and Zooming:* Simulate differences in patient size and organ scaling, reflecting inter-patient anatomical variability [8].

The paper by Cicek et al. [8] shows that data augmentation through rotations, scaling and elastic deformations improves the performance of volumetric segmentation tasks, even with limited training data.

3.4.2 Elastic Deformations

- *Elastic Transformations*: Mimic physiological movements and tissue deformations, such as breathing-induced organ shifts.

A paper proposed by Kamnitsas et al. [20] demonstrated the effectiveness of using elastic deformations and intensity variations in training deep models for brain lesion segmentation.

3.4.3 Intensity Transformations

- *Random Brightness & Contrast Adjustments:* Account for variations in imaging equipment calibration and patient-specific factors affecting image intensity [17].
- *Gaussian Noise Addition:* Reflects sensor noise and artifacts in medical images, enhancing robustness to such imperfections.

3.4.4 Spatial Transformations

- *Random Cropping & Padding:* Simulate variations in the field of view and ensure the model is not overfitting to specific spatial configurations [17].

3.4.5 Normalization

Normalization [8] is crucial for stabilizing and accelerating the training process. For each imaging modality, intensity normalization is performed using the following formula:

$$Normalized\ Image = \frac{Image - \min(Image)}{\max(Image) - \min(Image) + \epsilon} \tag{10}$$

where ϵ is a small constant (1×10^{-8}) to prevent division by zero. This step mitigates variations across different imaging protocols and scanners.

3.5 Handling Class Imbalance:

To address class imbalance, especially in tasks with small regions of interest, the following strategies are employed:

- Weighted Loss Functions [18]: Assign higher weights to underrepresented classes in the loss function.
- Oversampling: Increase the number of samples containing underrepresented classes.
- Focal Loss [41]: Utilize focal loss to focus learning on difficult examples.

4 Methodology

This section outlines the development of the proposed DUViT-Net model, including the inspiration behind its design, the detailed architecture, and how it builds upon and differs from the baseline DoubleU-Net. We also describe the data preprocessing, training procedures, and evaluation strategies to ensure robust and reproducible results. The primary motivation for developing DUViT-Net was to enhance medical image segmentation performance by effectively capturing local and global contextual information. While CNN-based architectures like U-Net [27] and DoubleU-Net[19] have demonstrated success in medical image segmentation, they are inherently limited by their local receptive fields, which may hinder their ability to model long-range dependencies critical in complex anatomical structures. Vision Transformers [9] have shown remarkable capabilities in capturing global context due to their self-attention mechanisms. By integrating ViT blocks into the DoubleU-Net architecture and combining their outputs with the encoder features, we aim to leverage the strengths of both CNNs and transformers to enhance segmentation performance.

4.1 Baseline Models: DoubeU-Net, nnU-Net, DiNTS, and MedVisionLlama

DoubleU-Net [19] employs two U-Net architectures sequentially to refine segmentation outputs (see Fig.1). The first U-Net performs coarse segmentation, then is refined by the second U-Net. The DoubleU-Net relies solely on convolutional operations, limiting its ability to capture global context.

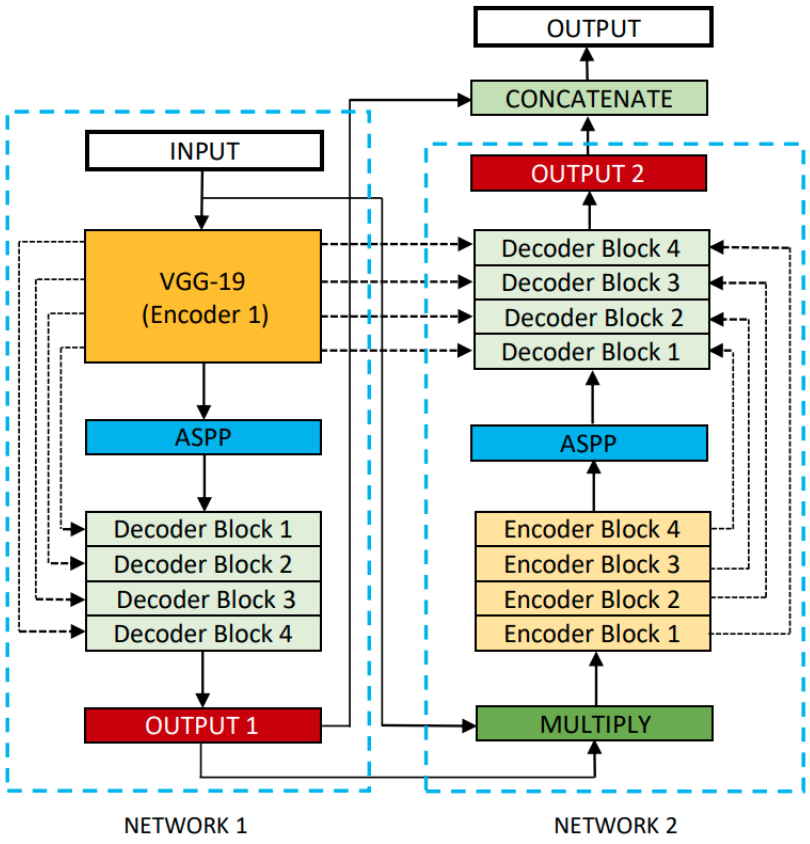

Figure 1: Block diagram of the original baseline DoubleU-Net with ASPP architecture proposed by [19] for medical image segmentation.

nnU-Net: Self-adapting Framework for U-Net-Based Medical Image Segmentation [16] stands out as a self-adapting framework that automates the configuration of U-Net architectures for various medical image segmentation tasks. The framework dynamically adjusts preprocessing steps, network architecture, and training procedures based on the specific characteristics of each dataset and task.

DiNTS: Differentiable Neural Network Topology Search for 3D Medical Image Segmentation [13] employ differentiable neural network topology search

within a 3D segmentation framework. The model automates the discovery of optimal network architectures tailored for specific segmentation tasks, allowing for dynamic adaptation based on the data's characteristics.

MedVisionLlama: Leveraging Pre-Trained Large Language Model Layers to Enhance Medical Image Segmentation [24] integrates pre-trained large language model (LLM) transformer blocks into ViT-based medical image segmentation models. The model leverages the pre-trained knowledge from extensive textual data to enhance feature representation by incorporating a frozen LLM transformer block into the encoder.

4.2 Proposed Model: DUViT-Net

The proposed *DUViT-Net* enhances the baseline DoubleU-Net by integrating ViT blocks after each encoder and combining their features with the encoder outputs to capture global contextual information. This integration allows the model to model long-range dependencies, complementing the local feature extraction capabilities of the CNN-based U-Nets. The architecture consists of two U-Net models, each augmented with ViT blocks, attention mechanisms, and SE blocks (see Fig. 2).

4.2.1 First U-Net(Coarse Segmentation):

The first U-Net depicted as U-Net1 + ViT in the Figure 2 consists of:

- *Encoder*: Utilizes a pretrained VGG19 [30] model modified to accept multi-channel input corresponding to the number of imaging modalities. The encoder extracts hierarchical features from the input image, progressively reducing spatial dimensions while increasing feature channels.
- *ViTBlock Integration:* A ViTBlock is incorporated after the encoder to capture global dependencies and contextual information that may be missed by convolutional layers alone. The ViTBlock processes the high-level feature maps, enabling the model to understand relationships between distant regions of the image.
- *Feature Fusion:* The features from the ViT block are projected to match the dimensionality of the encoder output using a convolutional layer. After upsampling the projected ViT features to match the spatial dimensions of the encoder output, they are added to the encoder features. This combination enriches the feature representation with local and global context before decoding.
- *ASPP Module*: Atrous Spatial Pyramid Pooling (ASPP) [6] modules are used to capture multi-scale information, enhancing the model's ability to handle objects of varying sizes. ASPP applies parallel atrous convolutions with different dilation rates, allowing the model to capture features at multiple scales without increasing computational complexity.

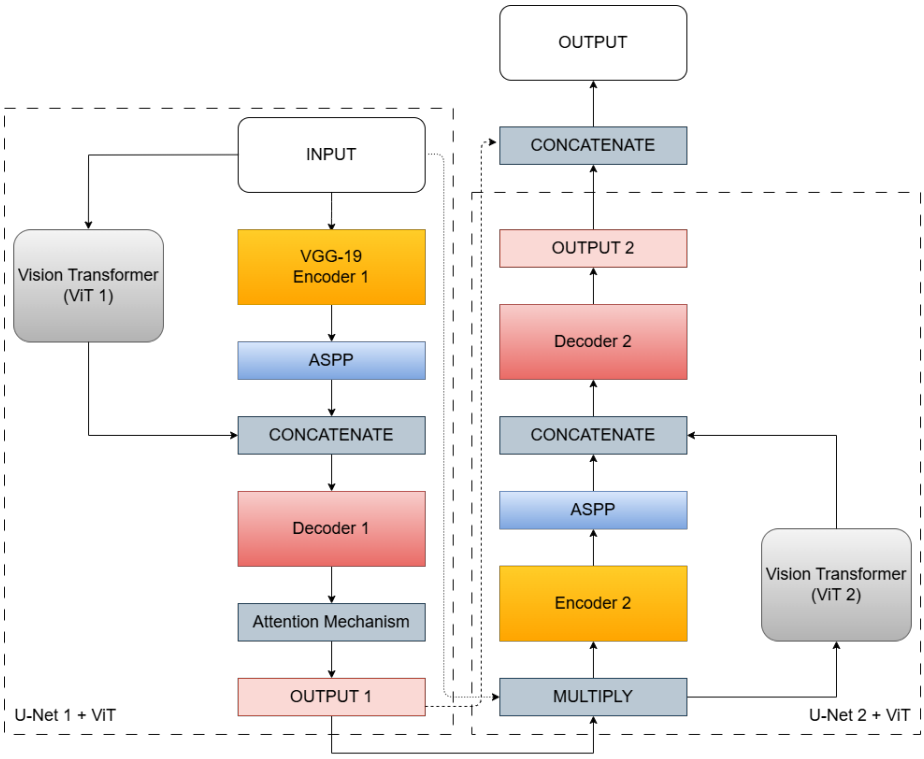

Figure 2: Proposed DUViT-Net architecture: A combination of Double U-net and Vision Transformers, enhanced with an Attention Mechanism.

- *Decoder:* Reconstructs the segmentation map by combining features from the encoder, ViTBlock, and ASPP modules through skip connections. The decoder progressively restores spatial dimensions, integrating multi-scale and global features to produce accurate segmentation masks.
- *Squeeze-and-Excitation(SE) Blocks:* Squeeze-and-Excitation blocks, introduced by [14], significantly enhance the representative capacity of convolutional neural networks by adaptively recalibrating channel-wise feature responses. The core idea behind SE blocks is to model the interdependencies between the channels of convolutional feature maps, allowing the network to emphasize informative features and suppress less useful ones selectively. This dynamic channel-wise feature recalibration is achieved through squeeze and excitation.
  - *Squeeze Operation:* The squeeze step aggregates global spatial information into a channel descriptor by performing global average pooling. For an input feature map $\mathbf{U} \in R^{C \times H \times W}$, where C is the number of channels, and H and W are the spatial dimensions, the squeeze op-

eration computes a channel-wise descriptor $\mathbf{z} \in R^C$ as follows:

$$z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i,j) \tag{11}$$

Here, $u_c(i,j)$ represents the activation at spatial location (i,j) in Channel c. This operation effectively captures each channel's global context by summarizing its spatial information.

– *Excitation Operation:* The excitation step captures channel-wise dependencies by passing the squeezed descriptors through a bottleneck of two fully connected (FC) layers. The first FC layer reduces the dimensionality by a factor of r (the reduction ratio), introducing non-linearity via a ReLU activation. The second FC layer restores the dimensionality to C and applies a sigmoid activation to obtain scaling factors $\mathbf{s} \in R^C$:

$$\mathbf{s} = \sigma\left(W_2 \cdot \delta\left(W_1 \cdot \mathbf{z}\right)\right) \tag{12}$$

Where:

* W_1 and W_2 are the weight matrices of the FC layers.
* δ denotes the ReLU activation function.
* σ represents the sigmoid activation function.

This mechanism allows the network to learn channel-wise dependencies and assign different levels of importance to each channel based on the global context.

4.2.2 Second U-Net(Refined Segmentation):

The second U-Net depicted as U-Net2 + ViT in the Figure 2 consists of:

- *Input Refinement*: The initial segmentation map from the first U-Net is used to refine the input, emphasizing relevant regions and providing additional contextual cues. This refinement helps the second U-Net focus on areas requiring higher precision.

- *Encoder*: A separate encoder processes the refined input, extracting further hierarchical features. This encoder is structurally similar to the first, ensuring consistency in feature extraction.

- *ViTBlock* Integration: Another ViTBlock is incorporated to capture the global context in the refined feature maps, further enhancing the model's ability to delineate complex structures accurately. The projected and upsampled ViT features are added to the encoder features, similar to the first U-Net. Combines both sets of features to enrich the representation before decoding.

- *Decoder*: Combining features from encoders and ViTBlocks through skip connections to produce the final, refined segmentation map. This decoder leverages both local and global features to achieve high segmentation accuracy.

4.2.3 ViTBlock Details:

- *Patch Embedding:* The feature maps are divided into non-overlapping patches, then flattened and linearly embedded to create a sequence of tokens suitable for the transformer. Positional embeddings are added to retain spatial information.
- *Transformer Encoder:* Multi-head self-attention layers model global relationships across the image patches, capturing dependencies between distant regions. The transformer encoder consists of alternating layers of multi-head self-attention and feed-forward networks, with layer normalization and residual connections.
- *Feature Reconstruction:* The transformed sequence is reshaped into spatial feature maps and integrated into the decoder, enriching the segmentation process with global context information. This integration ensures the decoder can access local and global features, facilitating accurate segmentation.

The inspiration for integrating ViT blocks into the DoubleU-Net architecture stems from the need to capture global contextual information that CNNs may miss due to their local receptive fields. By combining the features from the ViT blocks and the encoders, DUViT-Net can effectively capture both local and global contextual information. This fusion strategy enhances the feature representation, improving segmentation accuracy, especially in complex medical images. Incorporating attention mechanisms and SE blocks further refines the model's focus on relevant features, addressing the limitations of the baseline DoubleU-Net.

5 Dataset: Medical Segmentation Decathlon

Dataset Description: The Medical Segmentation Decathlon dataset (MSD) [3] is a comprehensive benchmark widely used for evaluating segmentation models across multiple medical imaging tasks. It comprises ten diverse challenges, each focusing on different anatomical structures and imaging modalities, including MRI and CT scans. The dataset is designed to test the generalization and robustness of segmentation models across various scenarios, making it an ideal platform for assessing model performance. This study focuses on tasks 01, 02, and 04 of the Medical Segmentation Decathlon dataset, which were chosen based on their clinical relevance, the diversity of segmentation challenges they present, and the practical limitations of computational resources.

- **Task 01 - Brain Tumor Segmentation:** Involves segmenting different tumour subregions from MRI scans.
- **Task 02 - Left Atrium Segmentation (Heart):** Entails segmenting the left atrium from cardiac MRI images, crucial for diagnosing and treating atrial fibrillation and other cardiac conditions.
- **Task 04 - Hippocampus Segmentation:** Involves segmenting the hippocampus from MRI scans, which is crucial for studying neurodegenerative diseases like Alzheimer's.

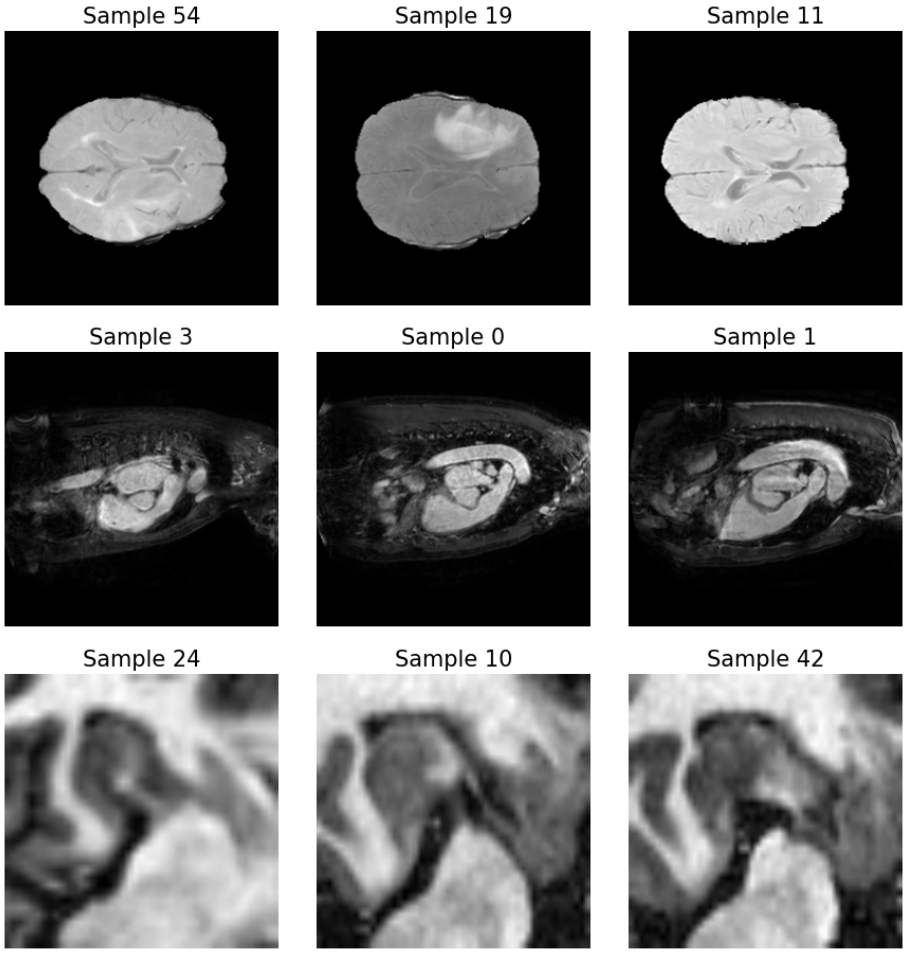

Figure 3: Visualization of random sample images from Task 01 (Brain Tumor), Task 02 (Left Atrium Heart), and Task 04 (Hippocampus) of the MSD Dataset.

Task	Modalities	No. of Images
Task01 BrainTumour	FLAIR, T1w, T1gd, T2w(MRI)	484
Task02 Heart	MRI	20
Task04 Hippocampus	MRI	260

Table 1: Summary of tasks, modalities, and number of images in the Medical Segmentation Decathlon Dataset.

Each task presents unique challenges, such as varying anatomical structures, imaging modalities, and segmentation complexities, providing a comprehensive evaluation of the models' capabilities.

	DataSplit		
Tasks	**Train**	**Validation**	**Test**
BrainTumour	339	97	48
Heart	14	4	2
Hippocampus	182	52	26

Table 2: The distribution of train, validation, and test datasets for Task 01 (brain tumour), Task 02 (heart), and Task 04 (hippocampus) from the Medical Segmentation Decathlon (MSD).

5.1 Data Pre-processing

Effective data preprocessing is essential for optimizing model performance and ensuring consistent input data quality. The following steps outline the preprocessing pipeline applied to the MSD dataset:

Slice Extraction: Volumetric images are sliced along the axial plane to obtain 2D images. Each slice is treated as an independent training, validation, and testing sample. This approach facilitates using 2D convolutions, reducing computational complexity compared to 3D models.

Normalization: Intensity normalization, as mentioned in Section 3.4.5, is performed per modality using z-score normalization, ensuring that the image intensities have zero mean and unit variance. This step mitigates variations across different imaging protocols and scanners, promoting consistent input data quality.

Data Augmentation: The augmentation techniques applied are specifically chosen to reflect plausible variations in medical imaging. The survey by [29] provides a comprehensive overview of image data augmentation techniques used

in deep learning, discussing their benefits, implementation methods, and best practices across various domains, including medical imaging. As detailed in Section 3.4 we applied the below data augmentation techniques,

- Geometric transformation [8] techniques such as horizontal flip, vertical flip, and random rotation was applied with probabilities of 50%.
- Elastic deformations [20] with Alpha parameter set to 120 and Sigma set to 6, applied with a 50% probability, and were applied on random images to simulate realistic anatomical changes.
- Intensity transformation [17] adjustments such as random brightness and contrast with 50% probability and Gaussian noise with 20% probability were applied. Multiple brightness and contrast adjustments were applied to diversify the dataset.
- Spatial transformation [8] parameters such as Shift Limit (0.0625), Scale Limit (0.1), and Rotation Limit (45 degrees) were used with a 50% probability.

These augmentations were applied on-the-fly, meaning each image was randomly transformed during each epoch, ensuring a rich and varied dataset. The probability of applying each transformation was set to introduce sufficient diversity in the training images while maintaining the plausibility of the augmented data for medical image segmentation tasks.

Handling Class Imbalance: To address the class imbalance, especially in tasks with small regions of interest, we applied strategies including weighted loss functions [18], oversampling, and utilizing focal loss [41].

6 Experimental Setup

This section details the experiments conducted to evaluate the proposed DUViT-Net model.

Experiments on the MSD Dataset: The experiments were conducted on Tasks 01 (Brain Tumor Segmentation), 02 (Left Atrium Segmentation), and 04 (Hippocampus Segmentation) of the Medical Segmentation Decathlon (MSD) dataset as described in section.5.

Baseline Comparison: To evaluate the effectiveness of the proposed DUViT-Net, we compared its performance with the baseline DoubleU-Net[19], reimplemented using the same preprocessing steps and training procedures for a fair comparison. Performance was assessed using the Dice coefficient, Jaccard index, precision, recall, F2 score, and HD95 distance to provide a comprehensive evaluation.

Comparison with MedVisionLlama: We also compared DUViT-Net with MedVisionLlama [24], a state-of-the-art model that integrates pre-trained Large Language Model (LLM) transformer blocks into ViT-based medical image segmentation models. We used the reported results from the MedVisionLlama paper for comparison. Both models were evaluated using the same tasks from the MSD dataset.

Ablation Study: An ablation study was conducted to assess the impact of various components in the proposed DUViT-Net. We tested different configurations based on our model's multiple components and with data augmentation to determine the contribution of each component to the overall performance. Below is the list of configurations we tested,

- **Baseline DoubleU-Net:** The standard DoubleU-Net architecture.
- **Without SE:** DoubleU-Net without the SE block.
- **With ASPP:** DoubleU-Net with the ASPP module.
- **With ViT**: DoubleU-Net with a single ViT block.
- **With 2ViTs:** DoubleU-Net with two ViT blocks (DUViT-Net).

Software and Libraries: The implementation was done using PyTorch, leveraging its flexibility and extensive support for deep learning operations. Additional libraries included Numpym Scipy, Nibabdel, Albumentationa, Timm and Scikit-learn.

Hardware Details: The experiments were executed on a workstation equipped with an NVIDIA Tesla T4 GPU and 16 GB of VRAM, which allowed for efficient data processing and model training.

Dataset Splits: The Medical Segmentation Decathlon dataset was divided into training (70%), validation (20%), and test (10%) sets. Stratified sampling ensured that each split maintained the distribution of classes and anatomical variations, preventing data leakage and ensuring fair evaluation of the models.

Reproducibility Measures: To ensure reproducibility, random seeds were set for all relevant libraries, including NumPy and PyTorch.

Hyperparameters: The proposed model DUViT-Net was trained using a learning rate of 1 $\times 10^{-4}$, a batch size of 4, and for 100 epochs. A learning rate scheduler ReduceLROnPlateau [2] was employed to reduce the learning rate by a factor of 0.1 if the validation loss did not improve for ten consecutive epochs. Adam optimizer [22] is chosen for its adaptive learning rate capabilities.

7 Results

7.1 Performance of DUViT-Net

The performance of the proposed DUViT-Net model across Tasks 01, 02, and 04 of the MSD dataset is summarised in Table.3.

The Dice coefficient is one of the most widely used evaluation metrics in medical image segmentation due to its effectiveness in quantifying the similarity between the predicted segmentation and the ground truth. Medical images often exhibit significant class imbalance; the region of interest (e.g., a tumour) may occupy a small portion of the image. The Dice coefficient is particularly suitable in such scenarios because it considers both false positives and false negatives equally, providing a balanced assessment of segmentation performance.

Model	**DUViT-Net**			
Metrics	**Task01-BrainTumour**	**Task02-Heart**	**Task04-Hippocampus**	**Avg.**
Dice	0.82	0.72	0.71	0.75
Precision	0.87	0.85	0.84	0.85
Recall	0.89	0.70	0.70	0.76
Jaccard	0.76	0.68	0.67	0.70
HD95	12.7	14.3	12.1	13.0
F2	0.89	0.73	0.72	0.78

Table 3: The performance of the proposed DUViT-Net model across various metrics on Task01 (brain tumour), Task02 (heart), and Task04 (hippocampus) from the Medical Segmentation Decathlon Dataset.

Brain Tumor Segmentation: DUViT-Net achieved a Dice coefficient of 0.82, indicating a high overlap between the predicted segmentation and ground truth. *Heart Segmentation:* Achieved a precision of 0.85, demonstrating the model's ability to identify positive pixels with minimal false positives correctly. *Hippocampus Segmentation:* Maintained consistent performance across all metrics, showcasing the model's robustness.

7.2 Comparison with State-of-the-Art

The performance of the DUViT-Net model is compared with the state-of-the-art models we discussed in the Section.2.4

These models nnU-Net [16], DiNTS [13], and MedVisionLlama [24] primarily report their performance using the Dice coefficient. These models are considered state-of-the-art in medical image segmentation. We also use the Dice coefficient as the primary evaluation metric to ensure a fair and direct comparison with these models. This allows us to benchmark the performance of our proposed DUViT-Net against nnU-Net, DiNTS and MedVisionLlama using the same criteria.

Model	Metrics (Task01_Brain_Tumour)					
	Dice	**Precision**	**Recall**	**Jaccard**	**HD95**	**F2**
nnU-Net [16]	0.78	-	-	-	-	-
DiNTS [13]	0.80	-	-	-	-	-
MedVisionLlama [24]	**0.84**	0.78	**0.91**	0.62	**11.2**	0.88
DUViT-Net(ours)	0.82	**0.87**	0.89	**0.76**	12.7	**0.89**

Table 4: Comparison of our DUViT-Net model performance with state-of-the-art models across various evaluation metrics on Task 01 Brain Tumour from the Medical Segmentation Decathlon dataset.

Table.4 shows that the proposed DUViT-Net model achieves competitive performance against the state-of-the-art models, demonstrating the effectiveness of integrating ViT blocks into the DoubleU-Net architecture. MedVisionLlama achieved the highest Dice score of 0.84, with DUViT-Net closely following at 0.82. DUViT-Net achieved a higher precision of 0.87 compared to MedVision-Llama's 0.78 and outperformed MedVisionLlama with a Jaccard index of 0.76 vs 0.62, indicating better overlap with ground truth.

7.3 Comparison with MedVisionLlama

A detailed comparison between the DUViT-Net and MedVisionLlama across Tasks 01 (Brain Tumour), 02 (Heart-Left Atrium), and 04 (Hippocampus) is presented in Table.5.

These results indicate that MedVisionLlama excels in specific metrics, particularly Dice score and recall. However, the proposed DUViT-Net offers competitive performance with higher Precision in some tasks and a better Jaccard index in all the tested tasks. This shows the model's ability to identify the overlap between ground truth and prediction.

Precision and Recall Trade-off: DUViT-Net consistently has higher Precision in all tasks. Higher Precision indicates that DUViT-Net produces fewer false positives, essential to reduce over-segmentation and avoid misclassifying healthy tissue as pathological. Meanwhile, MedVisionLlama shows higher recall across all tasks. Higher recall means that MedVisionLlama is better at detecting true positives and has fewer false negatives, which is crucial in medical diagnostics to ensure pathological regions are not missed.

F2 Score Interpretation: The F2 score emphasises recall, weighing it more heavily than Precision. MedVisionLlama achieves higher F2 scores in the Heart and Hippocampus tasks, aligning with its higher recall. In the Brain Tumor task, DUViT-Net slightly outperforms MedVisionLlama in the F2 score, indicating a balance between Precision and recall in this critical task.

Boundary Accuracy (HD95): MedVisionLlama has lower HD95 values across all tasks, indicating better boundary delineation. Lower HD95 means that the maximum deviation between the predicted and ground truth boundaries (ex-

Tasks	Metrics	MedVisionLlama	DUViT-Net
BrainTumour	Dice	**0.84**	0.82
	Precision	0.78	**0.87**
	Recall	**0.91**	0.89
	Jaccard	0.62	**0.76**
	HD95	**11.2**	12.7
	F2	0.88	**0.89**
Heart	Dice	**0.78**	0.72
	Precision	0.74	**0.85**
	Recall	**0.89**	0.70
	Jaccard	0.64	**0.68**
	HD95	**12.1**	14.3
	F2	**0.85**	0.73
Hippocampus	Dice	**0.72**	0.71
	Precision	0.72	**0.84**
	Recall	**0.88**	0.70
	Jaccard	0.65	**0.67**
	HD95	**11.6**	12.1
	F2	**0.84**	0.72

Table 5: Comparison of the performance of the MedVisionLlama model and the proposed DUViT-Net model across various metrics on Task01, Task02, and Task04 from the Medical Segmentation Decathlon Dataset.

cluding outliers) is smaller, which is vital for precise localisation in medical segmentation.

Overall, MedVisionLlama excels in Dice, Recall, F2 Score, and HD95. DUViT-Net outperforms the Jaccard index and Precision. Higher Precision and the Jaccard index suggest that it is effective at correctly identifying positive regions with fewer false positives. This is particularly beneficial when over-segmentation is a concern.

7.4 Ablation Studies:

An ablation study was conducted to evaluate the impact of different architectural components and training strategies on the performance of the DUViT-Net model for medical image segmentation. The study focused on Task 01 (Brain Tumor) from the Medical Segmentation Decathlon dataset. The configurations tested include the baseline DoubleU-Net and versions with the addition of Squeeze-and-Excitation (SE) blocks, Atrous Spatial Pyramid Pooling (ASPP) modules, single Vision Transformer (ViT) blocks, and dual ViT blocks (DUViT-Net). The models were evaluated with and without image augmenta-

tion to assess the influence of data diversity on performance. Table 6 presents the performance of each configuration. From the results, the following observations can be made:

Metrics	Image Agumentation	Ablation Study (Task01 BrainTumour) DoubleU-Net	with SE	with ASPP	with ViT	with 2ViT*
Dice	No	0.7296	0.7572	0.7690	0.7545	0.7708
	Yes	0.7374	0.7528	0.7485	0.7813	**0.7865**
Precision	No	0.7559	0.7702	0.7877	0.8256	0.8581
	Yes	0.7645	0.7768	0.7882	**0.8931**	0.8690
Recall	No	0.7527	0.7701	0.7620	0.7803	0.7884
	Yes	0.7631	0.7555	0.7618	0.8058	**0.8326**
Jaccard	No	0.7471	0.7593	0.7532	0.7540	0.7469
	Yes	0.7527	0.7468	0.7597	0.7446	**0.7601**
F2	No	0.7527	0.7694	0.7875	0.7872	0.7770
	Yes	0.7633	0.7592	0.7767	0.8016	**0.8226**

Table 6: Comparison of the various metric scores across different configurations of the proposed DUViT-Net model on the Task01 Brain Tumour from the MSD dataset. * refers to our proposed DUViT-Net model.

- *Impact of Image Augmentation:* Incorporating image augmentation techniques consistently improves performance across all configurations. With augmentation, the DUViT-Net configurations achieve the highest Dice score of 0.7865, indicating superior overlap between predicted segmentation and ground truth.
- *Impact of SE Block:* SE blocks enhance the model's ability to recalibrate channel-wise features, improving Dice scores. However, their impact is moderate, and benefits plateau with image augmentation.
- *Impact of ASPP Module:* From the Table.6 we can see that the ASPP module helps in effectively capturing multi-scale contextual information, leading to notable improvements in segmentation performance, especially in Dice and F2 scores.
- *Inclusion of ViT Blocks:*
  - *Single ViT Block:* Incorporating a single ViT block significantly increases Precision (0.8256 without augmentation and 0.8931 with augmentation), indicating a strong ability to identify tumour regions correctly. The high Precision with augmentation suggests that the model becomes more confident and accurate in its predictions.
  - *Double ViT Blocks (DUViT-Net):* DUViT-Net, with dual ViT blocks, exhibits the best balance between Precision and recall when combined with image augmentation. The high F2 score indicates exceptional performance in correctly identifying tumour regions (high recall) while maintaining a low false-positive rate (high Precision).

Overall, integrating ViT blocks significantly enhances the model's ability to capture global context, improving Precision and recall when combined with image augmentation. Augmentation consistently benefits all configurations, with the most substantial impact observed in models incorporating ViT blocks. DUViT-Net with image augmentation achieves the highest scores in critical metrics (Dice, Recall, F2 Score), demonstrating its effectiveness in medical image segmentation tasks. DUViT-Net offers an optimal balance, which is crucial in medical imaging where false positives and negatives have significant implications.

7.5 Visualisation of Segmentation Results

Figures.4, 5, 6 showcases the original images, ground truth masks, and segmentation outputs from our DUViT-Net model for the Task01_BrainTumour, Task02_Heart and Task04_Hippocampus from the MSD dataset.

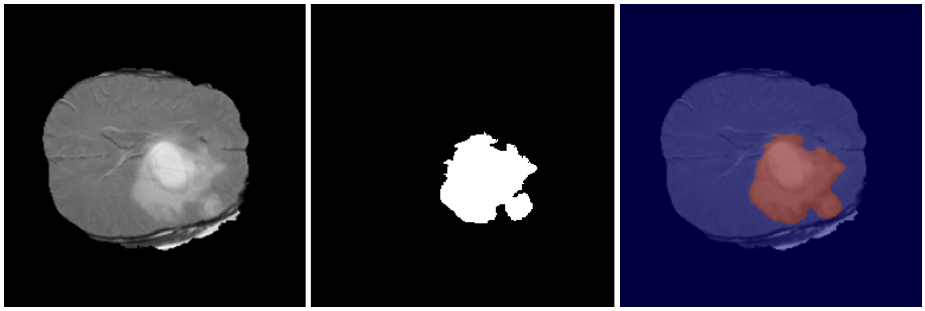

Figure 4: Comparison of the original images with ground truth and the prediction of our proposed DUViT-Net model on Task01 (BrainTumor) from the MSD Dataset

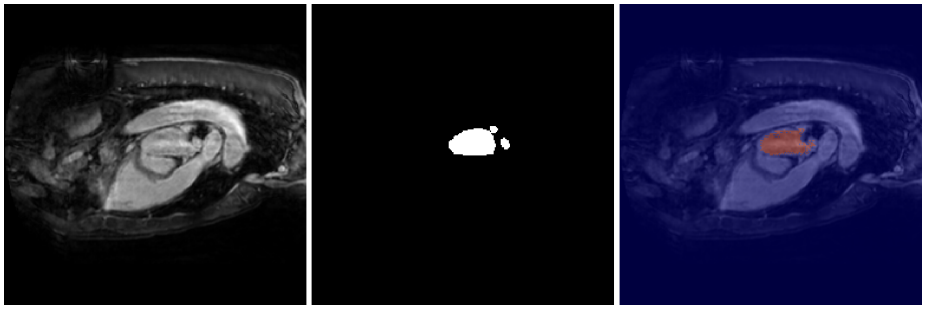

Figure 5: Comparison of the original images with ground truth and the prediction of our proposed DUViT-Net model on Task02 (Left Atrium Heart) from the MSD Dataset

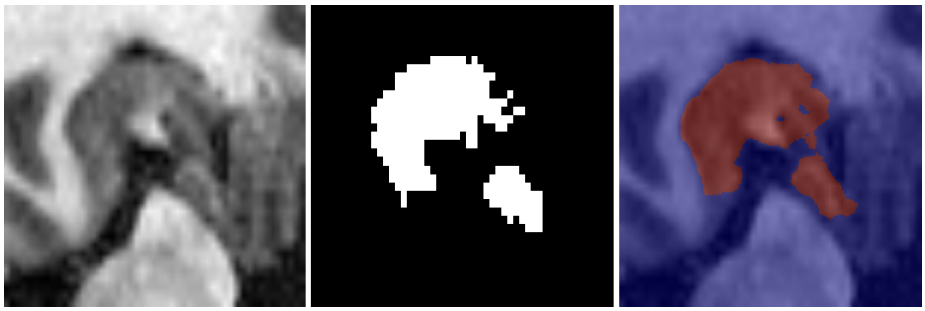

Figure 6: Comparison of the original images with ground truth and the prediction of our proposed DUViT-Net model on the Task04 (hippocampus) from the MSD Dataset

These visualisations (see Figures.4, 5, 6) highlight DUViT-Net's ability to accurately segment complex and irregularly shaped regions, demonstrating the effectiveness of integrating ViT blocks for capturing global context.

7.6 Training Performance of the DUViT-Net

The training and validation loss curves in Figure 7 show a steady decrease over epochs, indicating that the model is effectively learning from the data. The validation loss closely follows the training loss, suggesting the model generalises well without significant overfitting. Figure 7 illustrates the increase in the Dice

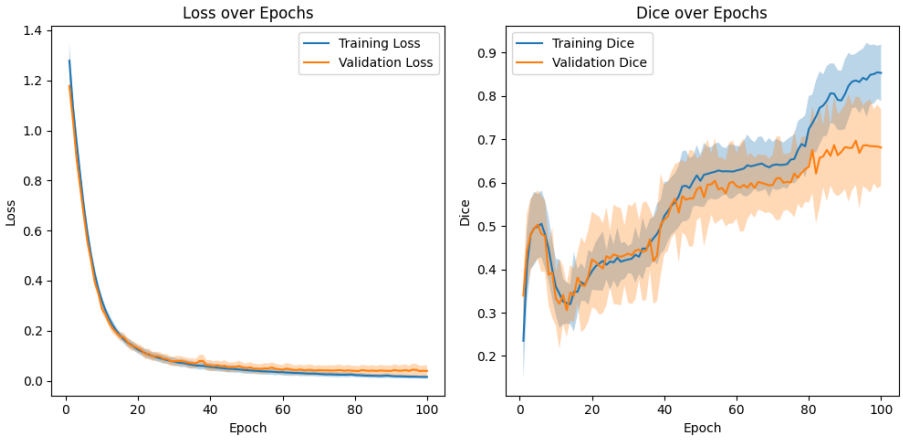

Figure 7: Training and validation loss and Dice curves from the proposed DUViT-Net model over 100 epochs for Task 01 Brain Tumour from the MSD Dataset.

coefficient for both training and validation sets over epochs. The curves show

that the model's segmentation accuracy improves consistently during training, reaching a plateau as it converges.

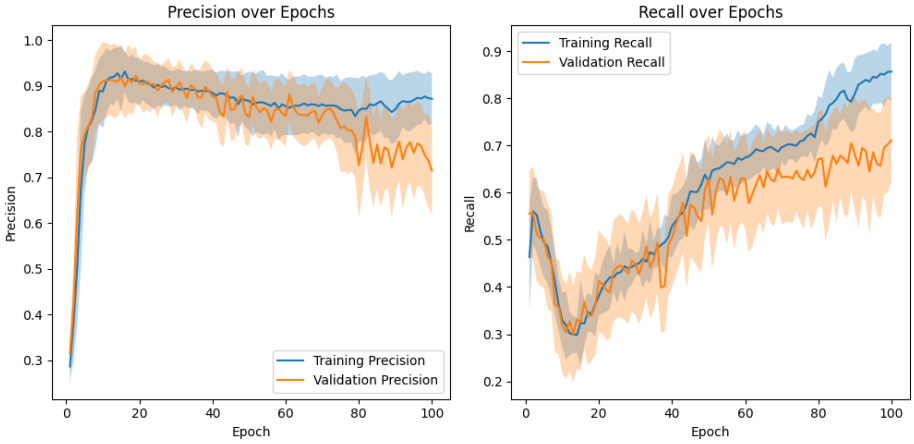

Figure 8: Training and validation Precision and Recall curves we got from the proposed DUViT-Net model over 100 epochs for Task 01 Brain Tumour from the MSD Dataset.

Figure 8 displays the Precision and recall metrics for both training and validation sets over epochs. The precision curve consistently increases, indicating that the model is improving its ability to correctly identify positive cases (i.e., correctly segmented regions). The recall curve also demonstrates improvement, reflecting the model's enhanced capability to detect all relevant instances of the target class.

The training and validation loss curves and the Dice coefficient, Precision, and recall plots demonstrate that the proposed model effectively learns the segmentation task. The faster convergence and lower loss values than the baseline model indicate that the ViT blocks contribute to a more efficient learning process.

Precision and Recall Analysis:

- *Precision Improvement:* The proposed model shows a higher precision throughout the training epochs than the baseline. This suggests the model is better at correctly identifying positive instances and reducing false positives.
- *Recall Enhancement:* The higher recall values indicate that the model captures all relevant positive instances more effectively, thus reducing false negatives.
- *Balance between Precision and Recall:* The simultaneous improvement in both Precision and recall demonstrates that the model is achieving a good

balance, which is crucial in medical image segmentation to ensure both accuracy and completeness of the segmented regions.

The minimal gap between the training and validation curves across all metrics suggests that the model generalises well to unseen data. This is crucial in medical image segmentation, where models must perform reliably on new patient data.

8 Discussion

8.1 Interpretation of Results

The proposed DUViT-Net model demonstrates significant improvements in segmentation performance across multiple tasks. The integration of ViT blocks enhances the model's ability to capture global context, complementing CNN's local feature extraction. This results in better delineation of complex anatomical structures and more accurate segmentation.

Ablation Studies: The ablation studies provide insights into the contribution of each component of the proposed model. Including the ASPP module improves the model's ability to capture multi-scale contextual information, enhancing the Dice score without data augmentation. However, with data augmentation, the improvement is less pronounced, suggesting that augmentation may provide sufficient variability for the model to learn multi-scale features. Incorporating ViT blocks significantly impacts the model's performance; adding a single ViT Block increases precision substantially, indicating that the model is highly confident in its positive predictions. However, the decrease in recall implies that some actual positive regions are missed. Adding a second ViT block helps to balance precision and recall. The model with two ViT blocks (DUViT-Net) and data augmentation achieves the highest Dice and F2 scores, demonstrating that this configuration effectively captures local and global features, improving overall segmentation performance. Also, the data augmentation consistently enhances performance across all configurations.

Comparison with State-of-the-Art: Compared to state-of-the-art models like nnU-Net, DiNTS, and MedVisionLlama, the proposed model achieves competitive performance, especially considering its 2D approach. Meanwhile, MedVisionLlama, which integrates LLM transformer blocks into ViT-based models, achieves higher Dice scores and precision in some tasks. However, in some instances, our model achieves a higher Jaccard index and recall, indicating better overlap with the ground truth and fewer missed regions. This suggests that while MedVisionLlama excels in maximizing overlap measures like Dice, the proposed model provides a balanced performance with strengths in different evaluation metrics, making it a viable alternative in settings where computational resources are limited.

8.2 Challenges and Limitations

- *Computational Resources:* Training transformer-based models can be computationally intensive due to the large number of parameters and the need to process extensive data. 3D models, especially when using transformer-based architectures, require significant GPU memory and processing power as they process volumetric data and compute attention across all voxels. However, the 2D approach used in this study mitigates these demands by processing 2D slices of medical images instead of 3D volumes. This reduces memory usage and computational complexity, enabling faster training with less GPU memory. The 2D model allows for more efficient processing, making it more feasible within resource-constrained environments compared to 3D models.
- *Dataset Size:* Some tasks in the MSD dataset have a limited number of images, which can affect the model's generalizability. Data augmentation helps but may not fully compensate for the lack of data diversity.
- *3D Context:* Processing 2D slices may result in the loss of some 3D contextual information inherent in volumetric data. Extending the model to handle 3D volumes could further enhance performance.

8.3 Recommendations for Future Work

Extending the model to process 3D volumes could capture spatial dependencies more effectively, potentially improving segmentation accuracy. Also, Experimenting with backbones like ResNet50 [12] and EfficientNet [33] may enhance performance by providing more robust feature extraction capabilities. Incorporating techniques like self-supervised pre-training or semi-supervised learning could improve results, especially with limited labelled data. Exploring the integration of pre-trained LLM transformer blocks, as done in MedVisionLlama, could potentially enhance the model's ability to capture complex data patterns.

9 Conclusion

This thesis demonstrates that integrating Vision Transformer blocks into a 2D DoubleU-Net architecture, resulting in DUViT-Net, can enhance medical image segmentation performance on the Medical Segmentation Decathlon dataset. DUViT-Net effectively captures local and global contextual information, improving accuracy and reliability. The model achieves competitive performance on multiple tasks from the MSD dataset, with significant improvements in metrics like the Jaccard index and precision compared to baseline and state-of-the-art models. These findings validate the potential of hybrid CNN transformer architectures for medical image segmentation, particularly in resource-constrained environments. The ablation studies highlight the significant impact of ViT blocks and data augmentation on performance. While models like MedVisionLlama leverage pre-trained LLMs for further enhancements, our approach shows

that substantial improvements can be achieved without relying on language models. The findings highlight the potential of combining CNNs and transformers in a unified framework, offering valuable insights for future research and applications in medical image analysis.

10 Contributions

Integration of ViTBlocks Enhances Performance: Introducing ViTBlocks into the DoubleU-Net framework enhances the model's capability to capture global context in 2D medical images, improving segmentation accuracy and boundary precision.

Comprehensive Evaluation Validates Effectiveness: The model is evaluated using a range of metrics, including the Dice Coefficient, Jaccard Index (IoU), Precision, Recall, F2 Score, and Hausdorff Distance (HD95), providing a detailed assessment of its performance across different aspects of segmentation quality. Using multiple evaluation metrics thoroughly assesses the model's strengths and areas for improvement.

Flexibility to Incorporate Other Backbones: While this study specifically used VGG-19 as the backbone for the model, the architecture is flexible and can easily be adapted to use other backbone architectures, such as ResNet or EfficientNet. However, it is important to note that this study did not include experiments with alternative backbones. Incorporating different backbones would offer flexibility to adapt the model to specific needs or constraints, depending on the task and computational resources.

Resource-Efficient Approach: Achieving high performance with a 2D model without extensive pre-training makes the approach practical for clinical settings with limited computational resources.

Potential of LLM Integration: The success of models like MedVisionLlama indicates that integrating pre-trained LLM transformer blocks could be a promising direction for future research

References

[1] Adams, R., and Bischof, L. Seeded region growing. *IEEE Transactions on pattern analysis and machine intelligence 16*, 6 (1994), pp. 641–647.

[2] Al-Kababji, A., Bensaali, F., and Dakua, S. P. Scheduling techniques for liver segmentation: Reducelronplateau vs onecyclelr. In *Intelligent Systems and Pattern Recognition* (Cham, 2022), A. Bennour, T. Ensari, Y. Kessentini, and S. Eom, Eds., Springer International Publishing, pp. 204–212.

[3] Antonelli, M., Reinke, A., Bakas, S., Farahani, K., Kopp-Schneider, A., Landman, B. A., Litjens, G., Menze, B., Ronneberger, O., Summers, R. M., et al. The medical segmentation decathlon. *Nature communications 13*, 1 (2022), pp. 4128.

[4] Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A. L., and Zhou, Y. Transunet: Transformers make strong encoders for medical image segmentation. *Medical Image Analysis* (2021).

[5] Chen, J., Mei, J., Li, X., Lu, Y., Yu, Q., Wei, Q., Luo, X., Xie, Y., Adeli, E., Wang, Y., et al. Transunet: Rethinking the u-net architecture design for medical image segmentation through the lens of transformers. *Medical Image Analysis* (2024), pp. 103280.

[6] Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence 40*, 4 (2017), pp. 834–848.

[7] Cheng, B., Schwing, A., and Kirillov, A. Per-pixel classification is not all you need for semantic segmentation. *Advances in neural information processing systems 34* (2021), pp. 17864–17875.

[8] Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., and Ronneberger, O. 3d u-net: Learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016* (Cham, 2016), S. Ourselin, L. Joskowicz, M. R. Sabuncu, G. Unal, and W. Wells, Eds., Springer International Publishing, pp. 424–432.

[9] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations* (2021).

[10] Guo, L., Fei, W., Dai, W., Li, C., Zou, J., and Xiong, H. Mixed-precision quantization of u-net for medical image segmentation. In *2022 IEEE International Symposium on Circuits and Systems (ISCAS)* (2022), pp. 2871–2875.

[11] Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H. R., and Xu, D. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI brainlesion workshop* (2021), Springer, pp. 272–284.

[12] He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), pp. 770–778.

[13] He, Y., Yang, D., Roth, H., Zhao, C., and Xu, D. Dints: Differentiable neural network topology search for 3d medical image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2021), pp. 5841–5850.

[14] Hu, J., Shen, L., and Sun, G. Squeeze-and-excitation networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018), pp. 7132–7141.

[15] Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2017), pp. 4700–4708.

[16] Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J., and Maier-Hein, K. H. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods 18*, 2 (2021), pp. 203–211.

[17] Isensee, F., Jäger, P., Wasserthal, J., Zimmerer, D., Petersen, J., Kohl, S., Schock, J., Klein, A., Ross, T., Wirkert, S., et al. batchgenerators—a python framework for data augmentation. *Zenodo 3632567* (2020).

[18] Jadon, S. A survey of loss functions for semantic segmentation. In *2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)* (2020), pp. 1–7.

[19] Jha, D., Riegler, M. A., Johansen, D., Halvorsen, P., and Johansen, H. D. Doubleu-net: A deep convolutional neural network for medical image segmentation. In *2020 IEEE 33rd International symposium on computer-based medical systems (CBMS)* (2020), IEEE, pp. 558–564.

[20] Kamnitsas, K., Ledig, C., Newcombe, V. F., Simpson, J. P., Kane, A. D., Menon, D. K., Rueckert, D., and Glocker, B. Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. *Medical image analysis 36* (2017), pp. 61–78.

[21] Kass, M., Witkin, A., and Terzopoulos, D. Snakes: Active contour models. *International journal of computer vision 1*, 4 (1988), pp. 321–331.

[22] Kingma, D. P., and Ba, J. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings* (2015), Y. Bengio and Y. LeCun, Eds.

[23] M. Beauchemin, K. T., and Edwards, G. On the hausdorff distance used for the evaluation of segmentation results. *Canadian Journal of Remote Sensing 24*, 1 (1998), pp. 3–8.

[24] Marthi Krishna Kumar, G., Chadha, A., Mendola, J., and Shmuel, A. Medvisionllama: Leveraging pre-trained large language model layers to enhance medical image segmentation, 2024.

[25] Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N. Y., Kainz, B., Glocker, B., and Rueckert, D. Attention u-net: Learning where to look for the pancreas. In *Medical Imaging with Deep Learning* (2018).

[26] Ostu, N. A threshold selection method from gray-level histograms. *IEEE Trans SMC 9* (1979), pp. 62.

[27] Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015* (2015), Springer, pp. 234–241.

[28] Schlemper, J., Oktay, O., Schaap, M., Heinrich, M., Kainz, B., Glocker, B., and Rueckert, D. Attention gated networks: Learning to leverage salient regions in medical images. *Medical Image Analysis 53* (2019), pp. 197–207.

[29] Shorten, C., and Khoshgoftaar, T. M. A survey on image data augmentation for deep learning. *Journal of big data 6*, 1 (2019), pp. 1–48.

[30] Simonyan, K., and Zisserman, A. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations (ICLR 2015)* (2015).

[31] Staib, L. H., and Duncan, J. S. Model-based deformable surface finding for medical images. *IEEE transactions on medical imaging 15*, 5 (1996).

[32] Taha, A. A., and Hanbury, A. Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool. *BMC medical imaging 15* (2015), pp. 1–28.

[33] Tan, M., and Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning* (2019), PMLR, pp. 6105–6114.

[34] Valanarasu, J. M. J., Oza, P., Hacihaliloglu, I., and Patel, V. M. Medical transformer: Gated axial-attention for medical image segmentation. In *Medical image computing and computer assisted intervention–MICCAI 2021* (2021), Springer, pp. 36–46.

[35] Valanarasu, J. M. J., Oza, P., Hacihaliloglu, I., and Patel, V. M. Medical transformer: Gated axial-attention for medical image segmentation. In *Medical image computing and computer assisted intervention–MICCAI 2021* (2021), Springer, pp. 36–46.

[36] Vaswani, A., Shazeer, N. M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. In *Neural Information Processing Systems* (2017).

[37] Wang, S., Peppa, M. V., Xiao, W., Maharjan, S. B., Joshi, S. P., and Mills, J. P. A second-order attention network for glacial lake segmentation from remotely sensed imagery. *ISPRS Journal of Photogrammetry and Remote Sensing 189* (2022), pp. 289–301.

[38] Wang, Z., Wang, E., and Zhu, Y. Image segmentation evaluation: a survey of methods. *Artificial Intelligence Review 53*, 8 (2020), pp. 5637–5674.

[39] Xie, Y., Zhang, J., Shen, C., and Xia, Y. Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021* (2021), Springer, pp. 171–180.

[40] Xu, G., Zhang, X., He, X., and Wu, X. Levit-unet: Make faster encoders with transformer for medical image segmentation. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)* (2023), Springer, pp. 42–53.

[41] Yeung, M., Sala, E., Schönlieb, C.-B., and Rundo, L. Unified focal loss: Generalising dice and cross entropy-based losses to handle class imbalanced medical image segmentation. *Computerized Medical Imaging and Graphics 95* (2022), pp. 102026.

[42] Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P. H., 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* (2021), pp. 6881–6890.

[43] Zhou, H.-Y., Guo, J., Zhang, Y., Yu, L., Wang, L., and Yu, Y. nnformer: Volumetric medical image segmentation via a 3d transformer. *IEEE Transactions on Image Processing 32* (2021), 4036–4045.