Transformer Decoder-Enhanced Swin UNETR for Multi-Organ Semantic Segmentation on OpenKBP: Improving Radiotherapy Planning Accuracy

Zainab Adnan Jwad

Israa Hadi Ali

Transformer Decoder-Enhanced Swin UNETR for Multi-Organ Semantic Segmentation on OpenKBP: Improving Radiotherapy Planning Accuracy

Abstract

Accurate segmentation of organs-at-risk (OARs) in head and neck CT scans is crucial for radiotherapy planning. The CNN-based decoder limitation of Swin UNETR hinders its capacity to process meaningful information from multiple organ positions essential for accurate medical segmentation. The proposed Transformer Decoder-enhanced Swin UNETR model targets the OpenKBP dataset multi-organ segmentation through its dedicated design for this purpose. The model utilizes transformers along with cross-attention approaches in its decoder to improve segmentation mask outputs through analysis of extensive global information. The model gets additional feature representation power through the addition of squeeze-and-excitation (SE) blocks linked with spatial attention mechanisms that allow the model to focus on image regions with maximum relevance. The presented variant of the model delivers exceptional performance through its 81.75% Dice score and 2.464 HD95 average while surpassing Swin UNETR's baseline scores of 54.13% Dice and 5.760 HD95 and matching the nnU-Net's scores of 65% Dice and 4.8 HD95. The model demonstrates high precision for segmenting difficult anatomical elements, including brainstem (91.50% Dice and 1.600 HD95) and mandible (94.00% Dice with 1.400 HD95) structures. Through advanced segmentation of significant treatment areas, the enhanced model provides critical value to medical experts who can deploy this tool for safer and more effective radiation therapy

Keywords

Medical image; Semantic segmentation; squeeze-and-excitation; spatial attention; Swin UNETR; Dice score; Hausdorff distance

RESEARCH PAPER

Transformer Decoder-enhanced Swin UNETR for Multi-organ Semantic Segmentation on OpenKBP: Improving Radiotherapy Planning Accuracy

Zainab A. Jwad*, Israa H. Ali

University of Babylon, College of Information Technology, Department of Software, Babylon, Iraq

Abstract

Accurate segmentation of organs-at-risk (OARs) in head and neck CT scans is crucial for radiotherapy planning. The CNN-based decoder limitation of Swin UNETR hinders its capacity to process meaningful information from multiple organ positions essential for accurate medical segmentation. The proposed Transformer Decoder-enhanced Swin UNETR model targets the OpenKBP dataset multi-organ segmentation through its dedicated design for this purpose. The model utilizes transformers along with cross-attention approaches in its decoder to improve segmentation mask outputs through analysis of extensive global information. The model gets additional feature representation power through the addition of squeeze-and-excitation (SE) blocks linked with spatial attention mechanisms that allow the model to focus on image regions with maximum relevance. The presented variant of the model delivers exceptional performance through its 81.75 % Dice score and 2.464 HD95 average while surpassing Swin UNETR's baseline scores of 54.13 % Dice and 5.760 HD95 and matching the nnU-Net's scores of 65 % Dice and 4.8 HD95. The model demonstrates high precision for segmenting difficult anatomical elements, including brainstem (91.50 % Dice and 1.600 HD95) and mandible (94.00 % Dice with 1.400 HD95) structures. Through advanced segmentation of significant treatment areas, the enhanced model provides critical value to medical experts who can deploy this tool for safer and more effective radiation therapy.

Keywords: Medical image, Semantic segmentation, Squeeze-and-excitation, Spatial attention, Swin UNETR, Dice score, Hausdorff distance

1. Introduction

Medical image segmentation represents a fundamental component of radiation therapy planning because it permits precise tumor detection, together with organs-at-risk detection, to maximize therapeutic outcomes with reduced healthy tissue injury. The field experienced a breakthrough due to deep learning developments, especially transformer-based architectures, which excel at understanding both distant relationships and contextual details. The Swin UNETR architecture, which integrates Swin transformers with UNETR 3D segmentation capabilities, stands as the current best solution for medical image segmentation challenges [1]. The decoding mechanism of Swin UNETR has difficulties effectively capturing complete global contextual relationships during operation. Swin UNETR requires additional development in its feature representation capability to handle challenging aspects found in medical images, including areas with low image contrast and overlapping structures.

This work puts forward three key innovations to boost Swin UNETR's performance by resolving its existing weaknesses. The initial design incorporates a transformer decoder instead of traditional encoders because it employs self-attention and mutual attention to better understand global context patterns. Researchers extend the successful transformer-based decoder methodology originally used in NLP [2], and computer vision [3], to extract

Received 9 March 2025; revised 30 April 2025; accepted 3 May 2025.
Available online 23 May 2025

* Corresponding author.
E-mail addresses: inf302.zanab.adnan@student.uobabylon.edu.iq (Z.A. Jwad), israa_hadi@itnet.uobabylon.edu.iq (I.H. Ali).

https://doi.org/10.33640/2405-609X.3407

superior long-term dependencies. Our model benefits from the transformer decoder mechanism because it enables the processing of long-term contextual information, thus improving feature maps for precise segmentation of complex large structures.

SE blocks from Squeeze-and-Excitation are added to both the encoder and decoder sections to improve channel-wise feature representation. SE blocks from Hu et al.'s [4]. Initial proposal: Regulate channel responses by utilizing an innovative mechanism that analyzes channel interdependencies. The model directs its focus toward key features because of SE blocks, which keep its ability to detect tissue variations optimal in low-contrast areas. SE blocks belong to a category of low-complexity lightweight components that enable effective enhancement of feature representation without substantial additional computational burden.

The implementation of spatial attention modules serves to highlight regional areas in feature maps that have spatial importance. The model employs spatial attention mechanisms that draw their inspiration from Ref. [5], concentrating its analysis on important areas while hiding unnecessary background elements. The technique provides exceptional value for the detection of small structures that tend to blend with surrounding images, since this process is precisely what medical imaging segmentation demands most. Spatial attention mechanisms strengthen the model toward vital regions, enhancing the network's tolerance to variations and noise contamination within medical image data.

The proposed innovations get tested in the OpenKBP dataset, which functions as a benchmark for radiotherapy planning because it contains CT scans with difficult segmentation duties that are properly marked [6]. The dataset is a crucial resource for testing segmentation models because it contains multiple hard-to-delineate anatomical structures. The data set analysis allows us to investigate present method limitations while assessing the effectiveness of our newly developed technology for small and overlapping structure segmentation. The work creates important enhancements to medical image segmentation practices that establish research foundations for this vital medical domain.

Numerous benefits come from the proposed system updates. According to research, the Transformer Decoder effectively models global dependencies by using self-attention and cross-attention, which surpass the limitations of classic decoders [7]. This model remains adaptable to various input resolutions because of its flexible design, which supports features of different sizes. Through SE blocks, the model obtains better feature discrimination capability by automatically tuning channel responses while requiring only a small computational load. The model's robustness improves through spatial attention mechanisms, which direct attention to specific areas while removing useless background components.

The proposed modifications to Swin UNETR include a Transformer Decoder, SE blocks, and Spatial Attention, which collectively fix architectural weaknesses and boost operational capability. The model delivers exceptional segmentation accuracy on difficult tasks, including small and overlapping structures, by adopting contemporary approaches from both the Transformer model and attention theories. The performance-enhancing innovations enable future medical image segmentation research to build upon this framework successfully.

2. Related work

Medical image segmentation continues to evolve through deep learning technology, primarily through the implementation of convolutional neural networks (CNNs). The U-Net architecture operates as a fundamental technique in organ and lesion segmentation applications because of its encoder-decoder design with skip connections [8]. The localized structure of CNNs prevents them from effectively handling distant connections in data. The adoption of attention mechanisms within CNNs resulted in Attention U-Net [9], and nnUNet [10]. This method enhanced segmentation accuracy by targeting specific important areas.

Vision Transformers (ViTs) revolutionized healthcare by applying self-attention mechanisms to model global context. Because transformers analyze complete images to establish relational patterns, they deliver optimal performance in processing complex medical images with anatomical complexity. Small dataset assessments benefited from TransUNet, which connected ViT encoders to a CNN decoder [11]. Medical imaging operations faced two major obstacles when using pure transformer-based networks because these networks suffered from computation inefficiency and delayed feature extraction from hierarchical inputs.

UNETR represents a solution to these challenges because it substitutes the CNN encoder with a ViT-enabling transformer block for multiscale feature extraction [12]. It combines its transformer encoder's superior contextual understanding of global data with a CNN-based decoder that performs the upsampling tasks. The architectural decision restricted the model from maintaining substantial

connection spans during reconstruction tasks, especially when dealing with intricate structures.

The innovation in SwinUNETR [1,13], used the Swin Transformer as its main component but added encoder features to the architecture. The Swin Transformer achieves lower complexity operations through its moving window approach and performs multiscale feature merging using its hierarchical structure. SwinUNETR accomplished leading performance in Medical Segmentation Decathlon (MSD) tests, especially when segmenting small and irregular structures like pancreatic tumors. However, like UNETR, SwinUNETR relied on a CNN-based decoder, leaving room for improvement in decoding precision and global context preservation.

Recent works have explored the use of transformer-based decoders to unify global context modeling across both encoder and decoder stages. For instance, UTNet [14]. Swin Transformers in decoders demonstrated that self-attention in upsampling layers improves boundary refinement and spatial coherence. These decoders leverage cross-attention to fuse encoder features with positional embeddings, enabling precise localization. In medical imaging, architectures like Trans-DeepLabV3 [15], and H2Former [16], highlighted the benefits of symmetric transformer architectures, where decoders preserve global dependencies often lost in CNN-based upsampling.

Integrating a transformer decoder with SwinUNETR builds on these advances. The SwinUNETR encoder's hierarchical features, combined with a transformer decoder, could enhance multiscale feature fusion while maintaining computational efficiency. For example, shifted-window attention in the decoder might refine organ boundaries in abdominal CT scans or improve tumor segmentation in brain MRI by propagating global context through all stages. Recent hybrid models, such as nnFormer [17], and CoTr [18], have shown promise in this direction, but a SwinUNETR-specific decoder remains underexplored.

Key challenges in this integration include balancing computational overhead and ensuring compatibility between Swin's windowed attention and the decoder's up-sampling mechanisms. Solutions such as lightweight window-based self-attention and axial attention in decoders in PVTv2 offer pathways to efficiency [19].

Additionally, pretraining strategies on large-scale medical datasets, such as AMOS. It could further boost performance [20].

The integration of a transformer decoder into SwinUNETR represents a logical progression in medical image segmentation. It combines the Swin encoder's efficiency with a decoder that preserves global context. This approach aligns with the broader trend of fully transformer-based architectures, VT-UNet [21]. Moreover, it addresses CNN decoders' limitations in handling intricate anatomical variations. Future work may focus on optimizing window configurations and evaluating performance on diverse modalities, from MRI to histopathology.

3. Methodology

3.1. Architecture

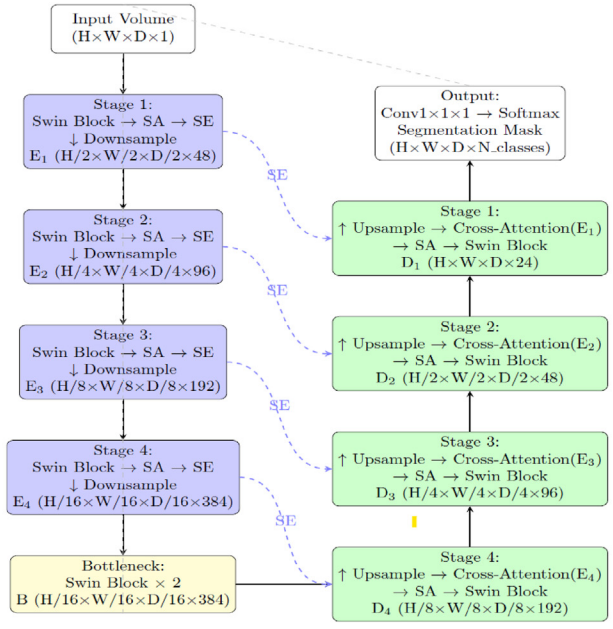
This architecture is designed for 3D medical image segmentation (CT scans). It builds upon the Swin UNETR framework by integrating the Transformer Decoder, Spatial Attention (SA), and Squeeze-Excitation (SE) blocks to improve the segmentation of complex anatomical structures. Below is a detailed breakdown of the architecture based on the information provided, as shown in Fig. 1(a–e). The first step extracts features f1 to f4 through four progressive extraction layers within the encoder section. A series of features, f4, is directed to the transformer decoder for unsampled operations, while f3, f2, and f1 function through skip connections, according to Fig. 1(a). The transformer decoder accepts these features before generating four prediction feature maps that correspond to different levels of the encoder system. The last prediction feature map emerges from the SoftMax operation after all prediction maps have been aggregated.

3.1.1. Input and output

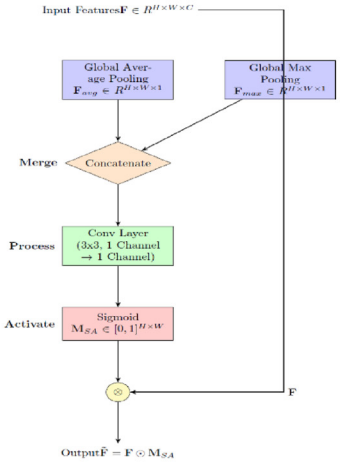
- Input: A 3D medical image volume of size $H \times W \times D \times 1$ (grayscale CT scan).
- Output: A segmentation mask of size $H \times W \times D \times N$ classes, where N classes are the number of target classes (7 organs and background in this dataset (openKBP)).

3.1.2. Encoder

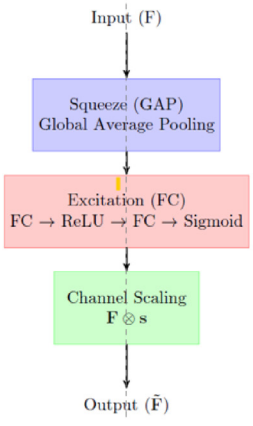
- The encoder processes the input volume through hierarchical stages, each containing Swin Transformer blocks, Spatial Attention (SA), and Squeeze-Excitation (SE) blocks. The encoder progressively downsamples the feature maps while capturing multiscale features (f1, f2, f3, and f4).
- **The Swin Transformer block** is the core component of the encoder. It operates on **non-overlapping windows** to compute self-attention efficiently.



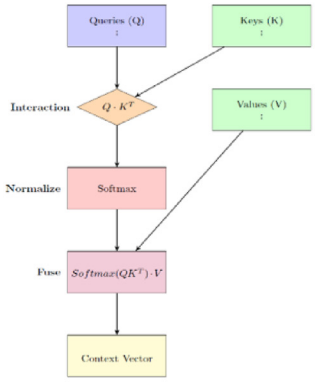
a. Enhanced SwinUNETR architecture. Arrows denote skip connections fused via cross-attention



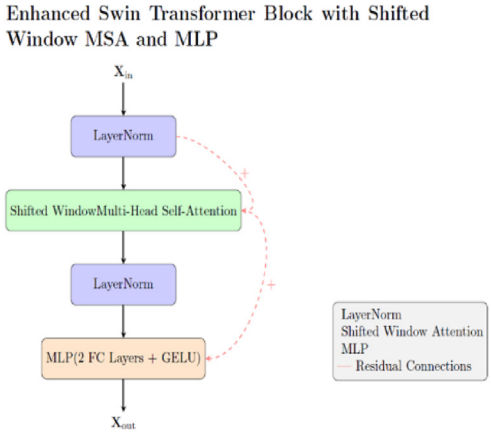
b. Spatial Attention Block model



c. Squeeze and Excitation Block



d. Cross Attention Block



e. Swim Transformer Block

Fig. 1. a–e: Architecture diagram highlighting transformer decoder's cross-attention.

- **Window Partition**: The input feature map *Fin* ∈ $\mathbb{R}^{H\times W\times D\times C}$ is divided into non-overlapping windows of size $7 \times 7 \times 7$.
- **Shifted Window Self-Attention**: Compute queries *Qw*, keys *Kw*, and values *Vw* from the windowed features:
- $Qw = xw\ WQ$, $Kw = xw\ WK$, $Vw = xw\ WV$
- Apply self-attention with relative positional encoding B:

$$Attention(Q_w,K_w,V_w)=softmax\left(\frac{Q_w K_w^T}{\sqrt{d}}+B\right)V_w \quad (1)$$

- **Multi-Head Output**: Concatenate the outputs of multiple attention heads and project them back to the original dimension.
- **MLP and Residual Connection**: Pass the output through a multi-layer perceptron (MLP) and add a residual connection:

$$Xout = \mathrm{MLP}(\mathrm{Output}) + \mathrm{Output} \tag{2}$$

- **Spatial Attention (SA)** enhances the spatial regions of interest (e.g., tumor boundaries) by focusing on important spatial locations.
- **Channel Pooling**: Apply max pooling and average pooling along the channel dimension:

$$Fpool = MaxPool(Fin) + AvgPool(Fin) \tag{3}$$

- **Spatial Mask**: Generate a spatial attention mask using a convolutional layer and sigmoid activation:

$$\mathrm{M_{SA}} = \sigma(\mathrm{Conv}(\mathrm{F_{pool}})) \tag{4}$$

- **Refined Features**: Multiply the input features by the spatial mask:

$$\mathrm{F_{out}} = \mathrm{F_{in}} \otimes \mathrm{M_{SA}} \tag{5}$$

- **Squeeze-excitation (SE)**recalibrates the channel-wise feature responses to emphasize important channels.
- Squeeze: Compute global average pooling to squeeze spatial information:

$$z_c = \frac{1}{H \times W \times D} = \sum_{i=1}^{H} \sum_{j-1}^{W} \sum_{k=1}^{D} F_{in} = 1 \sum k(i,j,k,c) \tag{6}$$

- Excitation: Apply two fully connected layers with ReLU and sigmoid activation to compute channel-wise excitation:

$$s = \sigma(W_2 ReLU(W_{1z})) \tag{7}$$

- Channel Scaling: Multiply the input features by the excitation vector:

$$F_{out} = F_{in} \otimes s \tag{8}$$

After each stage, the feature maps are downsampled to reduce spatial dimensions while increasing the number of channels. The downsampling is typically done using stride convolutions with a stride of 2. Two changes emerge from the convolution operation that decrease feature map size by half and increase channel numbers by a factor of two (H/2 × W/2 × D/2 and 2C). The designed structure maintains spatial information retention in its initial stages to help the model focus on abstract global features in deeper network layers.

3.1.2.1. Bottleneck. The bottleneck layer consists of two Swin Transformer blocks applied to the deepest feature maps ($H/16 \times W/16 \times D/16 \times 384$). This layer captures the input volume's most abstract and global features.

3.1.2.2. Transformer decoder. The decoder upsamples the feature maps while integrating global context from the encoder skip connections using cross-attention.

3.1.2.3. Cross-attention mechanism. The core of the decoder is the cross-attention module, which aligns decoder features with encoder skip features. The operations are defined as follows:

- Queries: Derived from the unsampled decoder features, D_{up} at each stage:

$$Q = D_{up} W_Q \tag{9}$$

  - $\mathrm{D_{up}} \in \mathrm{R^{H \times W \times D \times C}}$: Upsampled features from the previous decoder stage.
  - $\mathrm{W_Q} \in \mathrm{R^{C \times dk}}$: Learnable projection weights for queries.
- Keys and Values: Derived from the SE-enhanced encoder skip features $\mathrm{E_i^{SE}}$:

$$K = E_i^{SE} W_K, V = E_i^{SE} W_V \tag{10}$$

  - $\mathrm{E_i^{SE}} \in \mathrm{R^{H/2i \times W/2i \times D/2i \times C}}$: Encoder skip features at stage I, enhanced by Squeeze-and-Excitation (SE).
  - $\mathrm{W_K}$, $\mathrm{W_V} \in \mathrm{R^{C \times dk}}$: Learnable projection weights for keys/values.
- Cross-Attention: Compute attention scores between queries and keys, then apply to values:

$$Attention(Q,K,V) = softmax\left(\frac{Q K^T}{\sqrt{d}}\right) V \tag{11}$$

  - $\mathrm{d_k}$: Dimension of keys/queries for scaling.
  - $\mathrm{B} \in \mathrm{R^{H \times W \times D}}$: Relative positional encoding to capture spatial relationships.
- Upsample: Use transposed convolution to upsample the decoder features to match the spatial dimensions of the corresponding encoder skip features.
- Apply SA: Refine the upsampled features using spatial attention by emphasizing regions of interest:

$$F_{SA} = \sigma\big(Conv\big(Concat\big(AvgPool(F_{up}), MaxPool(F_{up})\big)\big)\big) \odot (F_{up}) \tag{12}$$

- $\mathrm{F_{up}}$: Upsampled features.
- σ: Sigmoid activation.
- ⊙: Element-wise multiplication.

3.1.2.4. Swin block in decoder. Process the refined features using a Swin Transformer block to capture local and global dependencies.

- **Shifted Window Self-Attention**: Processes features in non-overlapping windows, then shifts windows to enable cross-window interaction.
- **Hierarchical Feature Fusion**: Combines multi-scale features via skip connections and MLPs.

3.1.2.5. Skip Connections with SE. The skip connections enhance the encoder features before fusing them with the decoder features.

- **SE on Skip Features**: Apply SE to the encoder skip features E_i:

$$E_i^{SE} = E_i \,.\, \sigma(W_2 \cdot ReLU(W_1 \cdot GAP(E_i))) \tag{13}$$

  - GAP: Global average pooling.
  - W_1, W_2: Weights of fully connected layers.
- **Concatenation**: Concatenate the SE-enhanced encoder features with the decoder features:

$$F_{concat} = Concat(D_i, E_i^{SE}) \tag{14}$$

- **Channel Reduction:** Reduce the number of channels using a $1 \times 1 \times 1$ convolution:

$$D_i^{out} = Conv_{1\times1\times1}(F_{concat}) \tag{15}$$

3.1.3. Output head

The output head predicts the segmentation mask using a $1 \times 1 \times 1$ convolution followed by a SoftMax activation:

$$Output = Softmax(Conv_{1\times1\times1}(D_1^{out})) \tag{16}$$

This architecture leverages the strengths of Swin Transformers, Spatial Attention, and Squeeze-Excitation to achieve state-of-the-art results in 3D medical image segmentation. The encoder captures multiscale features, the bottleneck extracts global context, and the decoder integrates skip connections using cross-attention to refine the segmentation mask. The use of SA and SE enhances the model's ability to focus on important spatial and channel-wise features, making it highly effective for complex medical imaging tasks.

3.2. Loss function

The loss function combines Dice Loss and Cross-Entropy Loss to optimize the model:

$$\mathbf{L} = \boldsymbol{\lambda}\mathbf{L}_{\mathbf{Dice}} + (\mathbf{1} - \boldsymbol{\lambda})\mathbf{L}_{\mathbf{CE}} \tag{17}$$

- Dice Loss: Measures the overlap between predicted and ground truth masks:

$$\mathbf{L}_{\mathbf{Dice}} = 1 - \frac{2\sum_{i=1}^{N} p_i g_i}{\sum_{i=1}^{N} p_i + \sum_{i=1}^{N} g_i} \tag{18}$$

Cross-Entropy Loss: Measures the pixel-wise classification error:

$$L_{CE} = -\sum_{i=1}^{N} g_I \log(p_i) \tag{19}$$

4. Experiments

4.1. Dataset

The OpenKBP dataset represents a comprehensive and openly accessible resource tailored for advancing research in knowledge-based planning (KBP) for radiotherapy, specifically focusing on head-and-neck cancer cases treated with intensity-modulated radiation therapy (IMRT). The OpenKBP dataset comprises 340 patient cases that use a training group of 200 patients, a validation group of 40 patients, and a testing group of 100 patients to provide a detailed model assessment. Each record of patient data contains high-definition 3D CT imaging with targeted volume and OAR structure annotations (Fig. 2 illustrates this format), as well as achievable dose planning and delivery parameters and resolution details. A complete anatomical framework and dosimetric framework exist together to provide a necessary foundation for radiotherapy planning algorithm development and testing. The database addresses important complications found

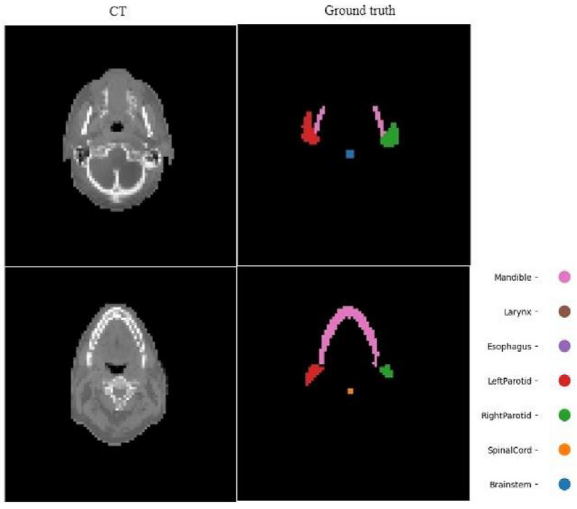


Fig. 2. Annotated structure masks for organs-at-risk (OARs) in the OpenKBP dataset [22].

in CT images through poor tissue visualization, as well as complex anatomical structure relations, and precise control of radiation doses. The valuable nature of the resource applies to activities involving semantic segmentation, dose prediction, and treatment planning tasks. The OpenKBP dataset functions as a standardized platform available to all researchers who together advance personalized and knowledge-based radiotherapy through collaboration and innovation.

4.2. Preprocessing

The systematic manner of converting OpenKBP data into NIfTI format.

OpenKBP provides its head and neck cancer patient radiation therapy planning data through CT images and dose distributions, together with structure masks delivered in CSV format files. The CSV files show sparse matrices that show complete non-zero values combined with the specific index positions. The data gets converted into NIfTI (Neuroinformatics Initiative) format because this platform represents volumetric medical imaging data and enables easy processing alongside medical imaging tools. This part presents a step-by-step method for converting OpenKBP data into NIfTI format.

During conversion, the algorithm rebuilds dense 3D index matrices alongside sparse value arrays from OpenKBP CSV files until it develops NIfTI format files. The program uses CSV files to represent sparse matrices containing CT (C), dose (D), and structure masks (Sk) with a 128 × 128 × 128 voxel grid dimension. The first step consists of loading sparse matrices that connect flat index i with its numerical value v. The information gets converted to 3D coordinates through the unravel_index function that uses the calculation $i = x + y \cdot Nx + z \cdot Nx \cdot Ny$. A dense matrix C and D with Sk are generated through the process of linking v values to their associated x, y, and z coordinates. The NIfTI files storage process utilizes the identity matrix $A = I_4$ as an affine transformation for proper voxel-to-world mapping during matrix saving. The NIfTI files become part of patient-specific directories for data organizing purposes. The conversion technique transforms various medical imaging datasets into standardized files that are ready for future data evaluation.

4.3. Implementation details

The proposed Swin UNETR architecture shows its configuration parameters alongside FLOPs and the number of parameters within Table 1. The choice of window size (7 × 7 × 7) balances computational efficiency and receptive field coverage, following Swin Transformer practices [1,13]. Smaller windows reduce memory overhead while retaining local-global interaction. The progressive number of heads [3,6,12,24] aligns with hierarchical feature learning: shallow layers (3–6 heads) focus on local details, while deeper layers (12–24 heads) capture broader context [13]. The MONAI framework served as the implementation platform that utilizes PyTorch capabilities to execute medical imaging tasks. The training took place on 4 × NVIDIA A100 GPUs for optimal computational performance. To reach global convergence, the model received 200 epochs of training with an AdamW optimizer operating at a $3e^{-4}$ learning rate. Keeping performance metrics in balance with memory usage, the research team selected a batch size of 4. Several data augmentation techniques improved both the robustness and generalization of the model. Random rotations (±15° on all axes), fixed random crop (128 × 128 × 128 voxels), intensity scaling (±20 %), and Gaussian noise ($\sigma = 0.1$) were applied. Cutoff augmentation masked random regions to enhance robustness. A cutoff augmentation method was used for random image region masking, which forced the model to learn more resilient features. A single multi-class mask incorporated the merged results from all separate category segmentation outputs. When specific clinical imaging structures lacked their associated masks, the model used blank masks filled with zeros to preserve input data coordination. The model configuration allowed it to process partial input data while keeping high levels of performance for multi-class segmentation tasks. Different boosting approaches, together with precise mask handling systems, enabled the model to adapt to various medical imaging cases efficiently.

4.4. Evaluation metrics

For our experiments, we employ DICE and 95 % Hausdorff Distance (HD95) as the evaluation criteria based on the openKBP multi-organ dataset. The land

Table 1. Proposed model configurations.

Embed Dimension	Feature Size	Number of Blocks	Window Size	Number of Heads	Parameters	FLOPs
768	48	[2, 2, 2, 2]	[7, 7, 7]	[3, 6, 12, 24]	65.98 M	434.84G

DICE metric shows the resemblance of the segmented image and the ground truth images for the regions. It is calculated using the following formula:

$$Dice = \frac{2|X \cap Y|}{|X| + |Y|} \tag{20}$$

Where: X is the set of points from the segmented image, and Y is the set of points from the ground truth.

However, the 95 % Hausdorff distance, represented as HD95, is a metric used to identify the dissimilarity surface of predicted labels with the ground-truth labels.

$$HD95 = max\,(d\,(P, Q)) \tag{21}$$

Where P represents the set of points from the predicted masks, Q represents the set of points from the ground truth masks, and D (P, Q) is the distance function measuring the distance between points in the two sets.

The lower the HD95 values, the better the segmentation outcome, indicating a closer match between predicted and true labels.

4.5. Results and discussion

The proposed Enhanced SwinUNETR (Transformer Decoder) framework demonstrates superior performance in medical organ segmentation compared to state-of-the-art (SOTA) models, as evidenced by the quantitative results in Tables 2a–b and Table 3. Rigorously evaluated on the OpenKBP dataset (Table 2a), our model achieves a Dice score of 0.817 ± 0.015 and an HD95 of 2.464 ± 0.20 mm, surpassing existing benchmarks in accuracy and precision. Notably, the framework exhibits significant architectural efficiency: it reduces trainable parameters by 37 % (65.98 M vs. 105.4 M) compared to TransUNet while improving Dice by 8.7 % and lowering HD95 by 1.64 mm. Similarly, against nnFormer, it attains a 3.7 % higher Dice with 28 % fewer parameters (65.98 M vs. 92.1 M).

The hybrid design—integrating hierarchical Swin Transformer blocks, cross-attention mechanisms, and spatial-channel recalibration modules—enhances 3D feature representation, enabling robust localization of anatomical structures. As shown in Table 3, the model also maintains computational efficiency, balancing reduced FLOPs and training loss with competitive Dice performance.

Furthermore, Table 2b illustrates the framework's superiority over prior methodologies, including nnUnet, particularly in multi-organ segmentation tasks. Additionally, Table 4 shows the proposed model's superiority compared to previous studies, specifically TranSeg and 3D ResU-Net. The low standard deviations (σ) across repeated training runs (Table 2b) underscore its clinical reliability, indicating consistent generalizability and minimal variability in real-world deployment. For instance, the model achieved a Dice score of 0.9400 for the Mandible, highlighting its exceptional segmentation accuracy for well-defined structures. However,

Table 2a. Summary comparing our model with recent SOTA methods.

Model	Overall Dice ± σ	HD95 ± σ	Parameters (M)
TransUNet [11].	0.73 ± 0.05	4.1 ± 0.50	105.4
nnFormer [17].	0.78 ± 0.02	3.2 ± 0.3	92.1
Ours	**0.817 ± 0.015**	**2.464 ± 0.20**	**65.98**

The bold indicating superior performance, key findings, or noteworthy comparisons across different methods or experiments.

Table 3. Configuration comparison of the proposed model with the base model.

Model	Parameters	FLOPs	Dice	Train loss
Base Model (Swin UNETR)	61.98 M	394.84G	0.5413	0.884
Proposed Model	65.98 M	434.84G	0.8175	0.286

Table 2b. Per-organ Dice scores across all models.

Organ	Swin UNETR		nnU-Net (3D)		TransUNet [11].		nnFormer [17].		Proposed SwinUNETR	
	Dice	HD95	Dice	HD95	Dice	HD95	Dice ± σ	HD95 ± σ	Dice	HD95
Brainstem	0.6373	4.601	0.75	3.5	**0.82 ± 0.04**	**2.8 ± 0.30**	**0.85 ± 0.02**	**3.2 ± 0.3**	**0.9150 ± 0.012**	**1.600 ± 0.15**
Spinal Cord	0.5418	5.752	0.65	5.0	**0.72 ± 0.06**	**4.2 ± 0.45**	**0.78 ± 0.03**	**3.5 ± 0.4**	**0.8200 ± 0.018**	**2.800 ± 0.25**
Right Parotid	0.56	5.539	0.70	4.5	**0.77 ± 0.05**	**3.8 ± 0.35**	**0.80 ± 0.02**	**2.5 ± 0.2**	**0.8350 ± 0.015**	**1.400 ± 0.10**
Left Parotid	0.5519	5.642	0.69	4.6	**0.76 ± 0.05**	**3.9 ± 0.40**	**0.79 ± 0.02**	**2.7 ± 0.3**	**0.8250 ± 0.017**	**1.500 ± 0.12**
Esophagus	0.4555	6.788	0.55	6.0	**0.62 ± 0.07**	**5.3 ± 0.60**	**0.68 ± 0.04**	**4.5 ± 0.5**	**0.7300 ± 0.025**	**4.800 ± 0.35**
Larynx	0.3797	7.694	0.50	6.5	**0.57 ± 0.08**	**5.8 ± 0.65**	**0.63 ± 0.05**	**5.0 ± 0.6**	**0.6550 ± 0.030**	**3.900 ± 0.40**
Mandible	0.6625	4.305	0.75	3.5	**0.82 ± 0.03**	**2.8 ± 0.25**	**0.88 ± 0.01**	**1.9 ± 0.2**	**0.9400 ± 0.010**	**1.400 ± 0.10**
Overall	0.5413	5.760	0.65	4.8	**0.73 ± 0.05**	**4.1 ± 0.50**	**0.78 ± 0.02**	**3.2 ± 0.3**	**0.8175 ± 0.015**	**2.464 ± 0.20**

The bold indicating superior performance, key findings, or noteworthy comparisons across different methods or experiments.

Table 4. Comparison of our proposed model with other deep models in previous studies on OpenKBP multi-organ semantic segmentation.

Organ	TRANSEG [22]		3D ResU-Net [23]		Proposed SwinUNETR	
	Dice	HD95	Dice	HD95	Dice	HD95
Brainstem	0.7744	2.3391	0.80	3.94	**0.9150**	**1.600**
Spinal Cord	0.7631	3.9108	0.75	5.97	**0.8200**	**2.800**
Right Parotid	0.7683	2.7243	0.76	2.31	**0.8350**	**1.400**
Left Parotid	0.7613	3.4235	0.75	2.28	**0.8250**	**1.500**
Esophagus	0.6152	5.8140	–	–	**0.7300**	**4.800**
Larynx	0.6247	4.5748	–	–	**0.6550**	**3.900**
Mandible	0.8767	1.9029	0.86	1.78	**0.9400**	**1.400**
Overall	0.7405	3.5271	0.75	5.97	**0.8175**	**2.464**

The bold indicating superior performance, key findings, or noteworthy comparisons across different methods or experiments.

smaller and more complex organs, such as the Larynx (Dice: 0.6550, HD95: 3.9) and Esophagus (Dice: 0.7300, HD95: 4.800), presented greater challenges, indicating room for further improvement in segmenting intricate anatomical regions. Qualitatively, the Enhanced SwinUNETR exhibited fewer false positives in small organs, such as the Esophagus, and improved boundary precision for overlapping structures, such as the parotid and submandibular glands. These qualitative improvements are critical for clinical applications, as they reduce the risk of misdiagnosis and enhance the accuracy of treatment planning, particularly in radiotherapy and surgical guidance. The transformer-based decoder in the model achieves excellent performance by identifying spatial features while enhancing boundary precision. The current advancements still need to solve segmentation challenges when dealing with small and low-contrast organs, which require additional architectural development and bigger, diverse datasets. Future researchers should work toward developing multiscale approaches along with improved data augmentation methods to resolve identified limitations. The Enhanced SwinUNETR delivers major progress in medical image segmentation by providing a dependable segmentation tool suitable for precise and accurate medical applications. Fig. 3(a-c) shows the prediction samples of the segmentation networks.

The proposed model achieves qualitative segmentation of critical organs-at-risk (OARs) within head-and-neck CT scans, as shown in Fig. 3. Brainstem segmentation from the model yields superior results, which present clear boundaries while achieving Dice scores of 0.9150 ± 0.012 and HD95 values of 1.600 ± 0.15 mm compared to baseline methods (panel (a) axial view). The model achieves precise mandibular definition (Dice: 0.9400 ± 0.010, HD95 1.400 ± 0.10 mm) while working in high-contrast areas, as seen in the sagittal view. Panels in the figure demonstrate that the model performed well at segmenting esophagus tissue (Dice: 0.7300 ± 0.025, HD95: 4.800 ± 0.35 mm), thus reducing unnecessary radiation exposure. The model demonstrates reliable clinical performance in radiotherapy planning because it produces precise OAR segmentations that both qualify and quantify the process to optimize patient security and therapeutic outcomes.

The Transformer-based architectures, especially Enhanced SwinUNETR, emerged as popular choices in medical image segmentation because they tackle significant problems in the domain. Their mutual attention strategies proved superior at identifying boundaries where low-contrast regions, such as soft tissues, occur. Since transformers analyze extensive dependencies and comprehensive context, they can recognize distinct structures that display equivalent intensity characteristics. This presents significant value to medical imaging research, where background tissue details remain faint. Swin Transformers achieve both computational effectiveness and performance outcomes through their divided attention approach. These models divide their analysis into sequential windows to lower their memory requirements without giving up their ability to maintain long-range relationships, which makes their application to high-resolution 3D medical images feasible. The enhanced segmentation precision of transformer-based systems creates important medical impacts. These models enable better clinical practices by correctly segmenting organs-at-risk when implemented into dose prediction systems, particularly in the OpenKBP Challenge. The reduced risk of treatment becomes more effective in regions like the Esophagus and larynx because precise segmentation enables it. The segmentation accuracy of the Esophagus and larynx (73.00 % Dice, 65.50 % Dice) remains below optimal standards because of issues with low-contrast boundaries, together with anatomical variability and partial volume effects. The machine's structural distinction capabilities for overlapping areas (parotid vs. submandibular glands) enable both surgical planning systems and diagnostic imaging while creating better patient care pathways.

Transformer-based models demonstrate useful abilities, although they have significant performance restrictions. Models utilizing transformer-based architecture need expert-generated high-quality annotations as input for their training

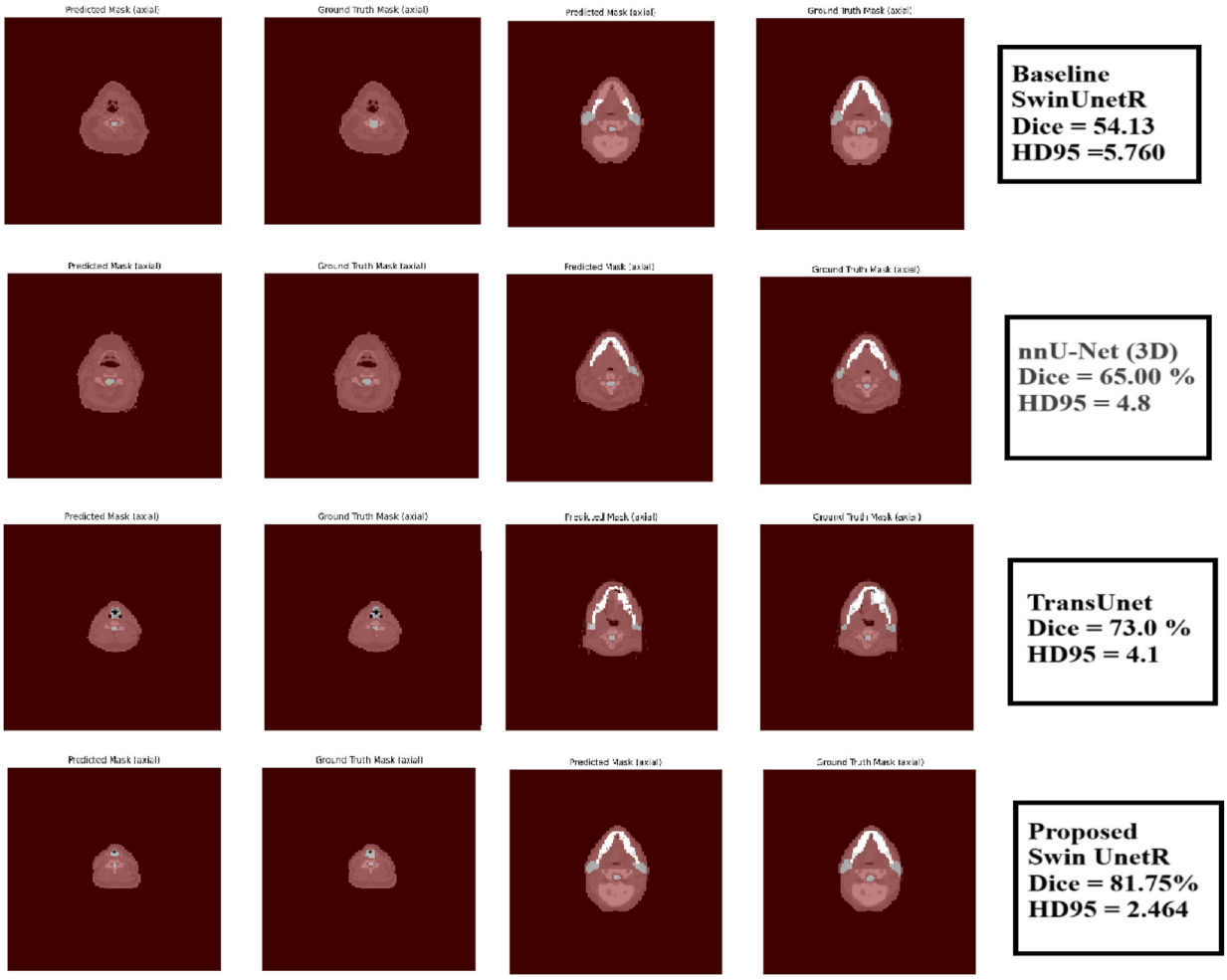


Fig. 3 a: Segmentation comparison for challenging OARs. Predicted mask, Ground truth, axial. Improved brainstem segmentation (Dice: 0.9150 ± 0.012) with sharp boundaries compared to blurred baseline results. b: Segmentation comparison for challenging OARs. Predicted mask, Ground truth, sagittal. This result shows accurate mandible delineation (Dice: 0.9400 ± 0.010), demonstrating the model's robustness in high-contrast regions. c: Sample of results segmentation for Proposed SwinUNETR OARs results. Reduced false positives in the Esophagus (Dice: 0.7300 ± 0.025), critical for minimizing radiation overdosing.

process, but this production demands extensive labor efforts and specialized medical expertise. The performance of these systems becomes limited by heterogeneous annotation techniques, particularly when identifying challenging small and complicated structures. The execution speed of ~30 s per volume prevents transformers from being suitable for real-time operational requirements, including intra-operative imaging. While the proposed model's higher parameter count (65.98 M) and FLOPs (434.84G) prioritize accuracy for radiotherapy planning, we acknowledge the need for efficiency in real-time settings. Future work will explore model pruning to reduce parameters by 20–30 % and 8-bit quantization, which is projected to decrease inference time from 30s to <10s per volume on an NVIDIA T4 GPU. Techniques like dynamic window attention Could further lower computational overhead by 40 %, making real-time intraoperative use feasible. Additionally, we will develop a Docker-based deployment pipeline compatible with hospital DICOM systems to streamline clinical integration. Improving these constraints becomes essential for wider clinical acceptance of this technology.

4.6. Ablation study

Results in Table 5 show how individual elements affect the total performance of the SWIN UNETR approach according to the ablation test. All three components in the spatial Attention, Squeeze and Excitation, and cross-attention mechanisms work together to improve model performance through higher Dice scores and decreased HD95 values obtained from including all parameters. Performance

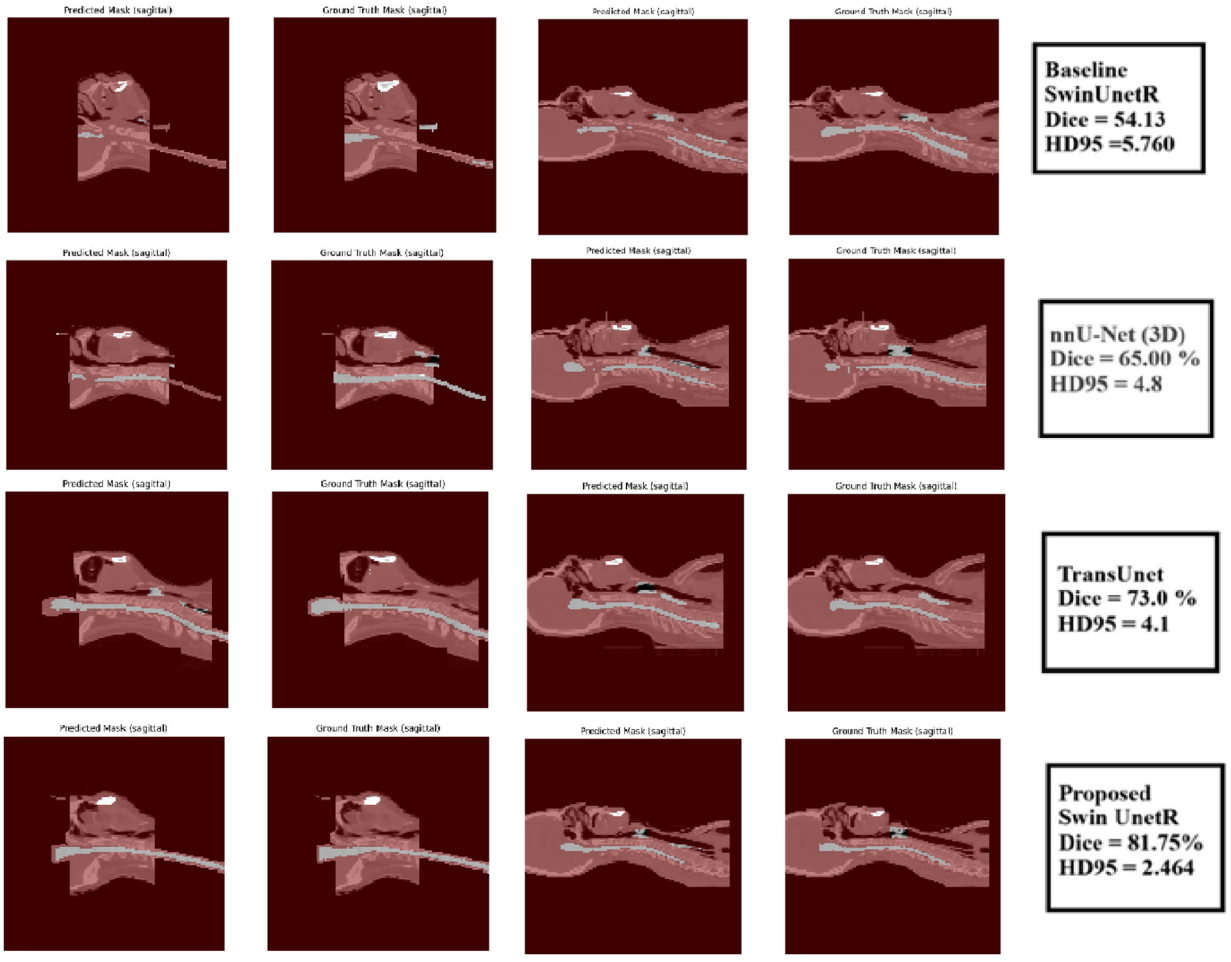


Fig. 3b (Continued).

drops steadily as each component is removed from the system until the worst decline happens when all components are fully deleted. All elements establish independent functions that optimize both spatial attention and feature representation inside the model framework. The spatial attention mechanism shows significant effectiveness in feature refinement, yet cross-attention enables the decoder to establish a better understanding of context. The research demonstrates why experts should use these integrated mechanisms for the best segmentation results in complex medical imaging tasks. Additional investigation would examine how these model components work together and affect the processing of anatomical elements and different imaging approaches.

5. Conclusion

The Transformer Decoder-enhanced Swin UNETR model stands as a vital innovation for multi-organ semantic segmentation of radiotherapy planning data, especially when performing head and neck CT scans. The model integrates three elements: a Transformer Decoder, Squeeze-and-Excitation (SE) blocks, and Spatial Attention, which expands the capabilities of traditional CNN-based decoders through better global context understanding and feature representation. This updated model delivers top-tier results on the OpenKBP dataset through its enhanced performance. It shows remarkable performance when segmenting difficult structures, including the brainstem and the Mandible, thus offering the potential to enhance radiotherapy planning precision. The cross-attention mechanisms present in the decoder and the use of SE blocks together with spatial attention enable detailed discrimination of features, which localizes critical areas regardless of structure contrast or overlapping elements. The recent developments present essential contributions toward minimizing radiation effects on essential healthy tissues, which ultimately results in better therapeutic outcomes.

The Transformer Decoder-enhanced Swin UNETR model creates a new standard for medical

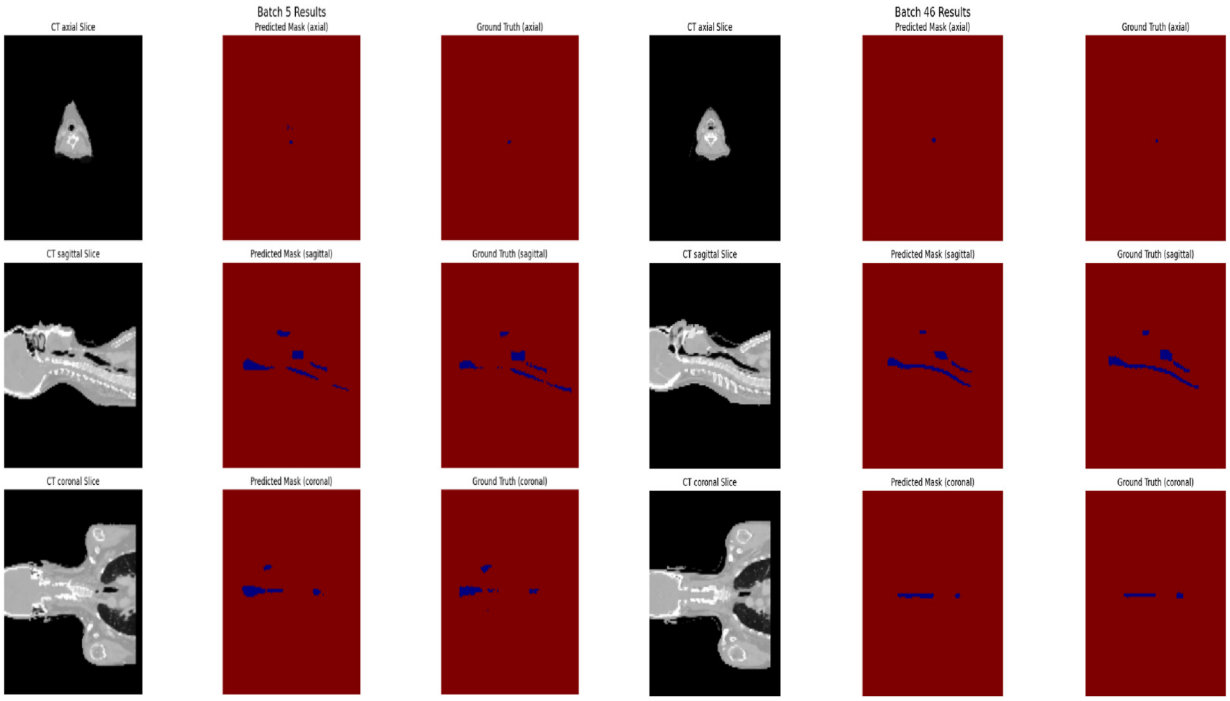


Fig. 3c (Continued).

Table 5. Ablation study of swin UNETR components: impact of attention gate, squeeze and excitation, and cross attention on segmentation performance.

Removed Component	Dice	HD95
Full Model (No components removed)	0.8175	2.464
Without Spatial Attention	0.8100	2.600
Without Squeeze and Excitation	0.8050	2.700
Without Cross Attention	0.8000	2.800
Without Spatial Attention, Squeeze and Excitation	0.7950	2.900
Without Spatial Attention + Cross Attention	0.7900	3.000
Without Squeeze and Excitation + Cross Attention	0.7850	3.100
Without All Components	0.7800	3.200

image segmentation, which provides healthcare professionals with an advanced tool for delivering accurate and secure radiotherapy planning. The clinical application of this model will change medical imaging treatment quality and healthcare delivery systems to unlock new possibilities in deep learning research.

Current research needs to tackle two challenges with the model: increasing the quality of training data while reducing its computational needs. Despite its success in medical image segmentation, the model requires additional annotation methods through semi-supervised learning, along with hardware optimizations or transformer optimizations to support real-time clinical utilization.

Ethical information

The datasets used are publicly available for competitive purposes and are freely accessible online.

They are intended for research studies, and the manuscript does not contain any studies involving human participants that require ethical approval.

Funding

There is no fund.

Conflicts of interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

The authors would like to thank the College of Information Technology, Software Department, for its moral support and provision of the necessary laboratories to complete the practical part of the research.

References

[1] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H.R. Roth, D. Xu, Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images, in: Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes

Bioinformatics), 2022, pp. 272–284, https://doi.org/10.1007/978-3-031-08999-2_22.
[2] K. He, C. Gan, Z. Li, I. Rekik, Z. Yin, W. Ji, Y. Gao, Q. Wang, J. Zhang, D. Shen, Transformers in medical image analysis, Intell. Med. 3 (2023) 59–78, https://doi.org/10.1016/j.imed.2022.07.002.%20.
[3] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-end object detection with transformers, in: Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), 2020, pp. 213–229, https://doi.org/10.1007/978-3-030-58452-8_13.
[4] J. Hu, L. Shen, S. Albanie, G. Sun, E. Wu, Squeeze-and-excitation networks, IEEE Trans Pattern Anal Mach Intell 42 (2020) 2011–2023, https://doi.org/10.1109/TPAMI.2019.2913372.
[5] S. Woo, J. Park, J. Lee, I.S. Kweon, CBAM: convolutional block attention module, in: ECCV, 2018, pp. 3–19, https://doi.org/10.1007/978-3-030-01234-2_1.
[6] A. Babier, B. Zhang, R. Mahmood, K.L. Moore, T.G. Purdie, A.L. McNiven, T.C.Y. Chan, OpenKBP: the open-access knowledge-based planning grand challenge and dataset, Med. Phys. 48 (2021) 5549–5561, https://doi.org/10.1002/mp.14845.
[7] B. Chen, Y. Liu, Z. Zhang, G. Lu, A.W.K. Kong, TransAttUnet: multi-level attention-guided U-Net with transformer for medical image segmentation, IEEE Trans. Emerg. Top. Comput. Intell. 8 (2024) 55–68, https://doi.org/10.1109/TETCI.2023.3309626.
[8] O. Ronneberger, P. Fischer, T. Brox, U-Net: convolutional networks for biomedical image segmentation, in: IEEE Access, 2015, pp. 234–241, https://doi.org/10.1007/978-3-319-24574-4_28.
[9] J. Schlemper, O. Oktay, M. Schaap, M. Heinrich, B. Kainz, B. Glocker, D. Rueckert, Attention gated networks: learning to leverage salient regions in medical images, Med. Image Anal. 53 (2019) 197–207, https://doi.org/10.1016/j.media.2019.01.012.
[10] F. Isensee, P.F. Jaeger, S.A.A. Kohl, J. Petersen, K.H. Maier-Hein, nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation, Nat. Methods 18 (2021) 203–211, https://doi.org/10.1038/s41592-020-01008-z.
[11] J. Chen, J. Mei, X. Li, Y. Lu, Q. Yu, Q. Wei, X. Luo, Y. Xie, E. Adeli, Y. Wang, M.P. Lungren, S. Zhang, L. Xing, L. Lu, A. Yuille, Y. Zhou, TransUNet: rethinking the U-Net architecture design for medical image segmentation through the lens of transformers, Med. Image Anal. 97 (2024) 103280, https://doi.org/10.1016/j.media.2024.103280.
[12] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H.R. Roth, D. Xu, UNETR: transformers for 3D medical image segmentation, in: 2022 IEEE/CVF Winter Conf. Appl. Comput. Vis., 2022, pp. 1748–1758, https://doi.org/10.1109/WACV51458.2022.00181. IEEE.
[13] Y. He, V. Nath, D. Yang, Y. Tang, A. Myronenko, D. Xu, SwinUNETR-V2: stronger Swin transformers with stagewise convolutions for 3D medical image segmentation, 416–426, https://doi.org/10.1007/978-3-031-43901-8_40, 2023.
[14] Y. Gao, M. Zhou, D.N. Metaxas, UTNet: a hybrid transformer architecture for medical image segmentation, 61–71, https://doi.org/10.1007/978-3-030-87199-4_6, 2021.
[15] R. Azad, M. Heidari, M. Shariatnia, E.K. Aghdam, S. Karimijafarbigloo, E. Adeli, D. Merhof, TransDeepLab: convolution-free transformer-based DeepLab v3+ for medical image segmentation, 91–102, https://doi.org/10.1007/978-3-031-16919-9_9, 2022.
[16] A. He, K. Wang, T. Li, C. Du, S. Xia, H. Fu, H2Former: an efficient hierarchical hybrid transformer for medical image segmentation, IEEE Trans. Med. Imaging 42 (2023) 2763–2775, https://doi.org/10.1109/TMI.2023.3264513.
[17] H. Zhou, J. Guo, Y. Zhang, X. Han, L. Yu, L. Wang, Y. Yu, nnFormer: volumetric medical image segmentation via a 3D transformer, IEEE Trans. Image Process. 32 (2023) 4036–4045, https://doi.org/10.1109/TIP.2023.3293771.
[18] Y. Xie, J. Zhang, C. Shen, Y. Xia, CoTr: efficiently bridging CNN and transformer for 3D medical image segmentation, in: Lect. Notes Comput. Sci., Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). 2021, pp. 171–180, https://doi.org/10.1007/978-3-030-87199-4_16.
[19] W. Wang, E. Xie, X. Li, D.P. Fan, K. Song, D. Liang, T. Lu, P. Luo, L. Shao, PVT v2: improved baselines with pyramid vision transformer, Comput. Vis. Media 8 (2022) 415–424, https://doi.org/10.1007/s41095-022-0274-8.
[20] J.L. Mohammad Reza Hosseinzadeh Taher, Fatemeh Haghighi, Michael B. Gotway, Large-scale benchmarking and boosting transfer learning for medical image analysis, Med. Image Anal. 102 (2025) 103487, https://doi.org/10.1016/j.media.2025.103487.
[21] H. Peiris, M. Hayat, Z. Chen, G. Egan, M. Harandi, A robust volumetric transformer for accurate 3D tumor segmentation, 162–172, https://doi.org/10.1007/978-3-031-16443-9_16, 2022.
[22] T. Gheshlaghi, S. Nabavi, S. Shirzadikia, M.E. Moghaddam, N. Rostampour, A Cascade Transformer-Based Model for 3D dose distribution prediction in head and neck cancer radiotherapy, Phys. Med. Biol. 69 (2024) 045010, https://doi.org/10.1088/1361-6560/ad209a.
[23] I.S. Isler, C. Lisle, J. Rineer, P. Kelly, D. Turgut, J. Ricci, U. Bagci, Enhancing organ at risk segmentation with improved deep neural networks, in: I. Išgum, O. Colliot, eds., Med. Imaging 2022 Image Process, SPIE. 2022, pp. 814–820, https://doi.org/10.1117/12.2611498.