Received: 9 May 2023 | Revised: 8 August 2023 | Accepted: 13 August 2023

DOI: 10.1002/mp.16750

RESEARCH ARTICLE

MEDICAL PHYSICS

A new architecture combining convolutional and transformer-based networks for automatic 3D multi-organ segmentation on CT images

Chengyin Li[1] | Hassan Bagher-Ebadian[2,3,4,5] | Rafi Ibn Sultan[1] | Mohamed Elshaikh[2] | Benjamin Movsas[2] | Dongxiao Zhu[1] | Indrin J. Chetty[2,6]

[1]College of Engineering - Dept. of Computer Science, Wayne State University, Detroit, Michigan, USA

[2]Department of Radiation Oncology, Henry Ford Cancer Institute, Detroit, Michigan, USA

[3]Department of Radiology, Michigan State University, East Lansing, Michigan, USA

[4]Department of Osteopathic Medicine, Michigan State University, East Lansing, Michigan, USA

[5]Department of Physics, Oakland University, Rochester, Michigan, USA

[6]Department of Radiation Oncology, Cedars Sinai Medical Center, Los Angeles, CA, USA

Correspondence
Indrin J. Chetty, Department of Radiation Oncology, Cedars-Sinai Medical Center, Los Angeles, CA, USA.
Email: indrin.chetty@cshs.org

Funding information
Varian Medical Systems, Siemens Healthineers

Abstract

Purpose: Deep learning-based networks have become increasingly popular in the field of medical image segmentation. The purpose of this research was to develop and optimize a new architecture for automatic segmentation of the prostate gland and normal organs in the pelvic, thoracic, and upper gastro-intestinal (GI) regions.

Methods: We developed an architecture which combines a shifted-window (Swin) transformer with a convolutional U-Net. The network includes a parallel encoder, a cross-fusion block, and a CNN-based decoder to extract local and global information and merge related features on the same scale. A skip connection is applied between the cross-fusion block and decoder to integrate low-level semantic features. Attention gates (AGs) are integrated within the CNN to suppress features in image background regions. Our network is termed "SwinAttUNet." We optimized the architecture for automatic image segmentation. Training datasets consisted of planning-CT datasets from 300 prostate cancer patients from an institutional database and 100 CT datasets from a publicly available dataset (CT-ORG). Images were linearly interpolated and resampled to a spatial resolution of (1.0 × 1.0× 1.5) mm^3. A volume patch (192 × 192 × 96) was used for training and inference, and the dataset was split into training (75%), validation (10%), and test (15%) cohorts. Data augmentation transforms were applied consisting of random flip, rotation, and intensity scaling. The loss function comprised Dice and cross-entropy equally weighted and summed. We evaluated Dice coefficients (DSC), 95th percentile Hausdorff Distances (HD95), and Average Surface Distances (ASD) between results of our network and ground truth data.

Results: SwinAttUNet, DSC values were 86.54 ± 1.21, 94.15 ± 1.17, and 87.15 ± 1.68% and HD95 values were 5.06 ± 1.42, 3.16 ± 0.93, and 5.54 ± 1.63 mm for the prostate, bladder, and rectum, respectively. Respective ASD values were 1.45 ± 0.57, 0.82 ± 0.12, and 1.42 ± 0.38 mm. For the lung, liver, kidneys and pelvic bones, respective DSC values were: 97.90 ± 0.80, 96.16 ± 0.76, 93.74 ± 2.25, and 89.31 ± 3.87%. Respective HD95 values were: 5.13 ± 4.11, 2.73 ± 1.19, 2.29 ± 1.47, and 5.31 ± 1.25 mm. Respective ASD values were: 1.88 ± 1.45, 1.78 ± 1.21, 0.71 ± 0.43, and 1.21 ± 1.11 mm. Our network outperformed several existing deep learning approaches using only attention-based convolutional or Transformer-based feature strategies, as detailed in the results section.

Conclusions: We have demonstrated that our new architecture combining Transformer- and convolution-based features is able to better learn the local and global context for automatic segmentation of multi-organ, CT-based anatomy.

KEYWORDS

3D segmentation, CT, deep convolutional network, multi-organ, transformer network

1 | INTRODUCTION

In the field of radiation therapy, precise targeting of tumor tissue while avoiding normal tissues is crucial for successful treatment.[1–3] One of the key steps in the planning process involves segmenting the treatment target and organs-at-risk (OARs) typically using planning CT images. Currently, the clinical practice for contour delineation involves a labor-intensive and operator-dependent manual process.[4–6] The manual contouring process in addition to often being inefficient can also suffer from inconsistencies in contouring preferences or related intra-and inter-observer uncertainties.[4,7] Inaccuracies in contouring impact on planning margin design—erroneous planning margins may lead to possible underdosage of the target and excess radiation delivered to surrounding healthy tissues.[8] To address these issues, a method for accurate automatic segmentation is needed to improve efficiency and consistency in radiation treatment planning.

Modern automatic multi-organ segmentation models can be roughly classified into two categories: conventional learning and deep learning-based segmentation.[3,9–11] In general, conventional learning-based approaches for building segmentation models have two major components[12]: (a) extraction of hand-crafted features to represent target organs, and (b) classification/regression model for segmentation. For instance, Glocker et al.[13] developed a supervised forest model that uses both class and structural information to jointly perform pixel classification and shape regression. To enhance the segmentation performance, Chen and Zheng[14] selected the most important features from the complete feature set using a hierarchical landmark detection method. Gao et al.[15] utilized multi-task random forests to segment prostate, bladder, rectum, and left and right femoral heads, jointly with a displacement regression task. Since these methods are typically created using low-dimensional hand-crafted features, their performance may be limited, particularly when the training datasets suffer from limited contrast impeding clear differentiation between organs at the boundaries, as is sometimes encountered with CT images.

Recently deep learning algorithms, which rely primarily on fully convolutional neural networks (CNN)-based U-net architectures[16–23] have been applied to the problem of organ segmentation for radiation treatment planning.[24–26] The U-Net is a popular architecture and comprises an encoder and decoder, where the encoder progressively reduces the resolution of CT scans to generate conceptual features across multiple scales. The decoder then reconstructs the extracted features for multi-organ segmentation. The U-net model incorporates skip-connections that combine the encoder and decoder outputs at different resolutions to maintain information lost during downsampling and improve performance. In pelvic organ segmentation, advanced U-net algorithms utilize supplementary techniques to facilitate the learning of more informative segmentation features. These techniques include a localization network for detecting the location of each organ prior to pixel-level segmentation,[27] a self-attention/Transformer mechanism for acquiring global features,[28] deep supervision for improving generality,[29] and multi-task learning strategies for capturing boundaries.[30]

While CNN-based U-Net's have demonstrated promise for medical image segmentation, they have limitations in modeling global context because the learning approach tends to be focused on local information.[31] To overcome this limitation, the vision Transformer (ViT)[32] has been proposed as an effective method to capture global dependencies and improve segmentation results for object structures with varying sizes and shapes. Studies have explored the integration of Transformers into U-Net architectures to enhance their performance in CT image segmentation. For instance, Chen et al.[33] used a Transformer between the encoder and decoder of U-Net to segment 2D abdominal CT scans and capture global context from U-Net feature maps. Similarly, Cao et al.[34] proposed a U-Net with a shifted-window (Swin) transformer (Swin-Unet) for 2D CT/MRI segmentation by replacing the convolutional blocks in U-Net with Swin Transformer blocks for both the encoder and decoder. More recently, UNETR[35] and SwinUNETR[36] were proposed for a multi-organ/multi-tumor segmentation approach on 3D CT scans. These networks replace the CNN-based encoder with a Transformer or Swin Transformer in the U-Net and have achieved state-of-the-art accuracy.[34–37] However, it is worth noting that while Transformer is effective at modeling global context, it is limited in capturing granular details due to a lack of spatial inductive bias in modeling local information, especially in the low data (high background) setting as is encountered with medical images.[38,39]

In this work, we developed and optimized a novel architecture, termed "SwinAttUNet" for 3D CT-based auto-segmentation of the prostate gland and surrounding OAR's, and other normal organs, including the lungs, liver, kidneys, and pelvic bones. SwinAttUNet bridges a 3D-U-Net and a Swin Transformer to take advantage of both architectures. SwinAttUNet includes a parallel encoder, a cross-fusion block, and a CNN-based decoder.

To our knowledge, this is the first network combining a 3D-based parallel CNN with a Transformer, along with several other unique features, for multiple organ segmentation. Details of the network architecture and quantitative evaluation of the model are presented.

2 | MATERIALS AND METHODS

2.1 | Data acquisition and preprocessing

All experiments were implemented on a server equipped with 8 Nvidia A100 GPUs, each with 40 GB of memory. All experiments were conducted in the PyTorch framework in Python 3.8.13, and each model was trained on a single GPU. Data augmentation was applied during training.

2.1.1 | Institutional dataset

Pelvic Multi-Organ Segmentation Dataset: Institutional review board (IRB) approval was obtained for this study. Planning CT and structure datasets for 300 prostate cancer patients were retrospectively selected. The 300 cases were randomly split into a training set of 225 cases, a validation set of 30 cases, and a testing set of 45 cases. The testing dataset was "held out" and therefore "unseen" relative to CT scans used for training and validation. All CT datasets were resampled into an isotropic voxel spacing of (1.0 × 1.0 × 1.5) mm,[40] and a Hounsfield unit (HU) range of [–50, 150] was used and normalized to [0, 1]. Subsequently, each CT dataset was cropped to a 192 × 192 × 64 voxel patch around the prostate/bladder/rectum regions, used in both training and inference for the 3D models. The models were trained for 200 epochs using the AdamW (Adaptive Moment Estimation Weighted, a variant of Adam where the weight decay is performed only after controlling the parameter-wise step size) optimizer with an initial learning rate of $5e^{-4}$. An exponential learning rate scheduler with a warmup of 5 epochs was applied to the optimizer. Random flip, rotation, and intensity scaling were used as augmentation transforms, with probabilities of 0.1, 0.1, and 0.2, respectively. The training datasets were increased by a factor of approximately 175 using data augmentation. The training process for 200 epochs required approximately 16.5 h.

Ground-truth segments were available for all image datasets consisting of physician drawn contours for the prostate gland (target) and surrounding normal tissues (bladder and rectum). The automatic contours generated by our network were compared to those of the ground-truth contours to evaluate the performance of the network.

2.1.2 | Public dataset

CT Organ Segmentation Dataset (CT-ORG): A publicly available dataset (CT-ORG)[41] was used for training and evaluation of our network for auto-segmentation of other organs. Details of the CT-ORG dataset are provided by Rister et al.[41] The dataset consisted of 100 CT scans, each of which included manual (ground-truth) contours of the lungs, liver, bladder, kidney, and pelvic bones. The first 19 CT datasets were "held out" and used solely for testing. The remaining 81 datasets were used for training following the process of Rister et al.[41] Each CT dataset was resampled with a voxel size of (2 × 2 × 5) mm, and input patches of size 128 × 128 × 64 were applied. Each CT dataset was truncated to HU range of [−1000, 1000] and normalized [−1, 1] over this range. The same augmentation and training strategy as the institutional dataset was applied to the CT-ORG dataset. The training process for 200 epochs required approximately 10.5 h.

2.2 | Network architecture

As depicted in Figure 1, we introduce the SwinAttUNet, which is a 3D network and is trained using 3D CT image datasets. SwinAttUNet includes a parallel encoder, a cross-fusion block, and a CNN-based decoder. The parallel encoder consists of a CNN branch (CB) and a Transformer branch (TB), which independently extracts local details and global contextual information. The cross-fusion block merges local and global features on the same scale. The CNN-based decoder is designed to adapt the fused information and thereby improve model stability while maintaining performance. A skip connection is applied between the cross-fusion block and decoder to integrate low-level semantic features. Attention gates (AGs) are integrated within the CNN to suppress features in image background regions, and focus attention to important regions of the image (targets and OAR's). All convolution blocks are 3D convolutions with a kernel size of three, and all transformer blocks are Swin Transformers (with window-based self-attention and shifted-window-based self-attention). We use two blocks for both convolution and transformer operations.

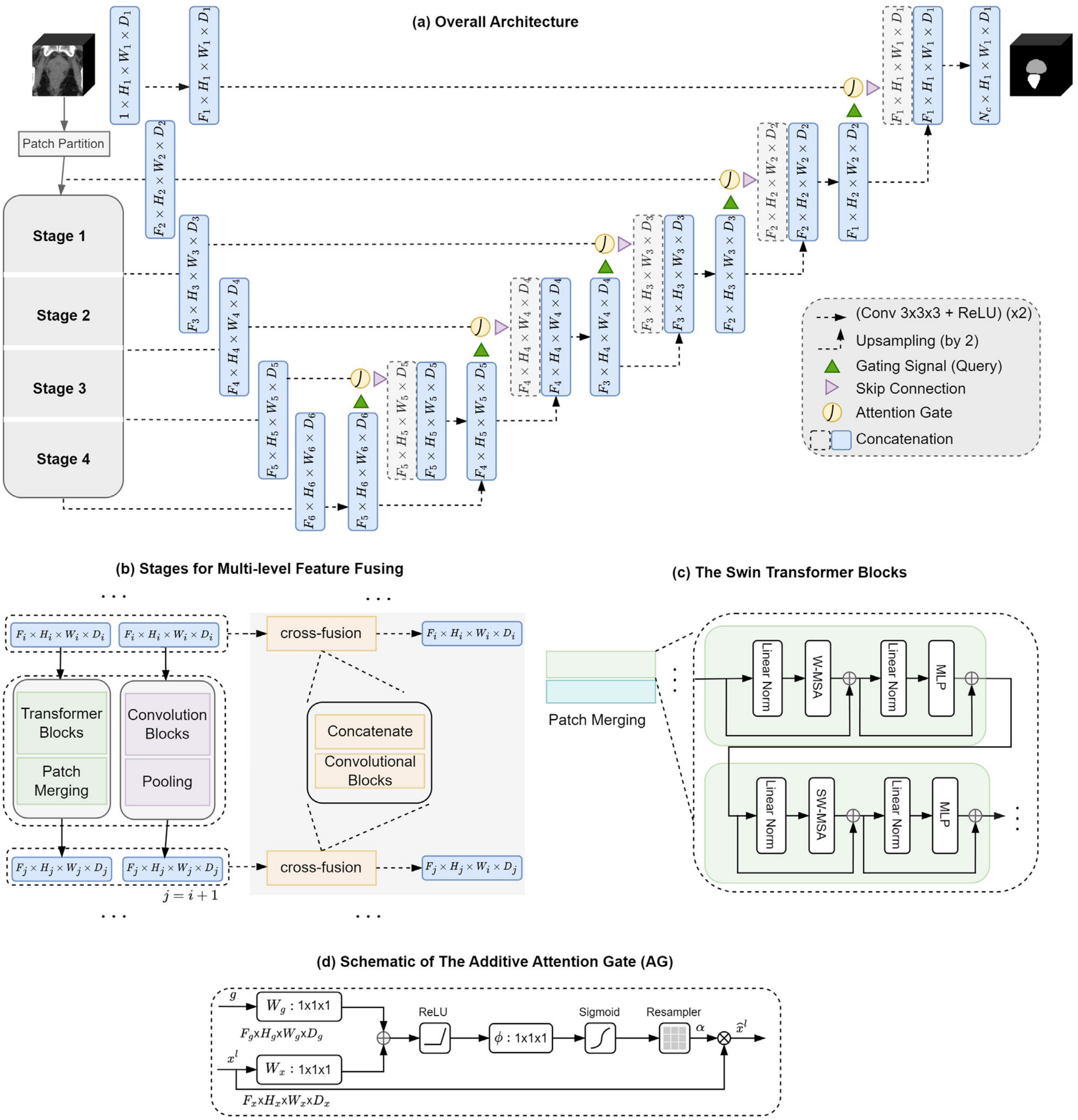
(a) Overall Architecture
Patch Partition
Stage 1
Stage 2
Stage 3
Stage 4
$1 \times H_1 \times W_1 \times D_1$
$F_1 \times H_1 \times W_1 \times D_1$
$F_2 \times H_2 \times W_2 \times D_2$
$F_3 \times H_3 \times W_3 \times D_3$
$F_4 \times H_4 \times W_4 \times D_4$
$F_5 \times H_5 \times W_5 \times D_5$
$F_6 \times H_6 \times W_6 \times D_6$
$F_4 \times H_5 \times W_5 \times D_5$
$F_3 \times H_4 \times W_4 \times D_4$
$F_2 \times H_3 \times W_3 \times D_3$
$F_1 \times H_2 \times W_2 \times D_2$
$N_c \times H_1 \times W_1 \times D_1$
(Conv 3x3x3 + ReLU) (x2)
Upsampling (by 2)
Gating Signal (Query)
Skip Connection
Attention Gate
Concatenation
(b) Stages for Multi-level Feature Fusing
$F_i \times H_i \times W_i \times D_i$
Transformer Blocks
Patch Merging
Convolution Blocks
Pooling
$F_j \times H_j \times W_j \times D_j$
$j = i + 1$
cross-fusion
Concatenate
Convolutional Blocks
(c) The Swin Transformer Blocks
Patch Merging
Linear Norm
W-MSA
MLP
SW-MSA
(d) Schematic of The Additive Attention Gate (AG)
g
W_g : 1x1x1
$F_g x H_g x W_g x D_g$
x^l
W_x : 1x1x1
$F_x x H_x x W_x x D_x$
ReLU
ϕ : 1x1x1
Sigmoid
Resampler
α
$\hat{x}^l$

FIGURE 1 (a) The architecture of SwinAttUNet for pelvic segmentation with 3D CT inputs. (b) Parallel CNN and Transformer blocks for encoder with a cross-fusion module. (c) The architecture of two successive Swin Transformer Blocks, W-MSA, and SW-MSA are multi-head self-attention modules with regular and shifted windowing configurations, respectively. (d) Schematic of the proposed additive AG. Input features (x^l) are scaled with attention coefficients (α) computed in AG. Spatial regions are selected by analyzing both the activations and contextual information provided by the gating signal (g) which is collected from a coarser scale. Grid resampling of attention coefficients is done using trilinear interpolation. AG, attention gate.

2.2.1 | Swin transformer branch for 3D inputs

Our SwinAttUNet architecture features a multi-scale design which enables generation of hierarchical feature maps at different stages.[35,36] As illustrated in Figure 1, the encoder takes as input an image volume $X \in R^{H \times W \times D \times C}$, where H, W, and D represent the spatial height, width, and depth, respectively, and C (1 for CT images) is the number of channels. A 3D token with a patch resolution of (H', W', D') has a dimension of $H' \times W' \times D' \times C$. The patch partitioning layer creates a sequence of 3D tokens with size $\frac{H}{H'} \times \frac{W}{W'} \times \frac{D}{D'}$ that are projected into a C'-dimensional space via an embedding layer. To efficiently model token interactions, we partition the input volumes into non-overlapping windows and compute local self-attention within each region. Specifically, at layer l, we use a window of size $M \times M \times M$ to evenly divide a 3D token into $\left\lceil \frac{H'}{M} \right\rceil \times \left\lceil \frac{W'}{M} \right\rceil \times \left\lceil \frac{D'}{M} \right\rceil$ windows. The encoder block outputs in layers l and $l + 1$ are computed as shown in Figure 1c, where W-MSA and SW-MSA denote regular and shifted window partitioning multi-head self-attention modules, respectively. A 3D cyclic shifting is also adopted for efficient batch computation of shifted windowing.[36,42]

2.2.2 | CNN branch for 3D inputs

Our CNN encoder branch composes of a series of convolutional layers with a skip connection to improve network stability. The use of convolutional layers in the encoder helps to detect local patterns and features such as edges and corners in the image. Specifically, it first applies a convolutional layer with 36 (1 × 1 × 1) spatial filters with stride 1 to the input data, and then passes it through four down-sampling residual blocks. Each residual block consists of one tri-linearly down-sampled layer followed by two 3D convolutional layers. The first convolutional layer has a 1 × 1 × 1 spatial filter with stride 1 in each direction, while the second convolutional layer uses 3 × 3 × 3 filters with the same stride. A skip connection used in ResNet[43] is applied between the outputs of the first and second convolutional blocks.

2.2.3 | Cross-fusion for two branches

To fully utilize both local and global features in our encoder, we use a parallel structure with a CNN and transformer blocks at each stage. To fuse these features, we introduce a cross-fusion module (shown in Figure 1b). This module takes two inputs with the same shape, $F_i \times H_i \times W_i \times D_i$, for the i-th stage, where F_i is the channel size. The module concatenates these two inputs and passes them through two layers of 3 × 3 × 3 convolution with residual connections. The output of this module is a fused feature map with the same shape as the input, which is then used as input for the proceeding decoding operations. This simple and efficient module allows us to combine the strengths of both CNN and transformer blocks in our encoder.

2.2.4 | Attention-enabled decoder

Standard CNN architectures gradually down-sample the feature-map grid to capture a large receptive field and semantic contextual information. However, reducing false positive predictions for small, variably shaped objects remains challenging. To address this issue, existing segmentation frameworks rely on separate object localization models. Here, we propose integrating AGs into a standard CNN model[29] to achieve the same objective without training multiple models or adding parameters. Unlike localization models in multi-stage CNNs, AGs progressively suppress features in irrelevant background regions without the need to crop regions of interest between networks.

Additive AGs are employed to modulate feature responses through skip connections, to determine a gating vector for each pixel enabling focus on relevant regions at each multiscale level. Although more computationally intensive than multiplicative attention, previous studies[29] have shown that additive AGs achieve superior predictive accuracy. An additive vector concatenation-based attention method[44] was adapted, in which the output of the n^{th} multi-scale encoding convolutional block (x^l) was added to the output of the $(n + 1)^{th}$ multi-scale decoding convolutional block (x^g), and the *ReLu* activation function applied to the combined activations. The input undergoes a channel-wise 1 × 1 × 1 convolutional layer, batch normalization layer, and sigmoidal activation layer is then multiplied and concatenated to the input of the n^{th} multi-scale level decoding convolutional block. Figure 1(d) illustrates the attention gating mechanism.

2.3 | Performance evaluation

2.3.1 | Loss functions

We utilize cross-entropy loss

$$\mathcal{L}_{CE} = \frac{1}{C} \sum_{k=1}^{i} y_k \log \hat{y}_k \tag{1}$$

and Dice loss

$$\mathcal{L}_{Dice} = \frac{1}{C} \sum_{k=1}^{C} \left(1 - \frac{2 \sum_{i \in I} y_k^i \hat{y}_k^i}{\sum_{i \in I} y_k^i + \sum_{i \in I} \hat{y}_k^i} \right) \tag{2}$$

for training, where C is the number of classes, I represents the whole image, y_k and $\hat{y}_k$ are the ground truth mask and the predicted segmentation from the model of class k, respectively. The overall loss function was cast as an equally weighted summation:

$$\mathcal{L}_{total} = \mathcal{L}_{CE} + \mathcal{L}_{Dice} \tag{3}$$

2.3.2 | Evaluation metrics

Dice score and 95% Hausdorff Distance (HD95) were used to evaluate the accuracy of segmentation in our

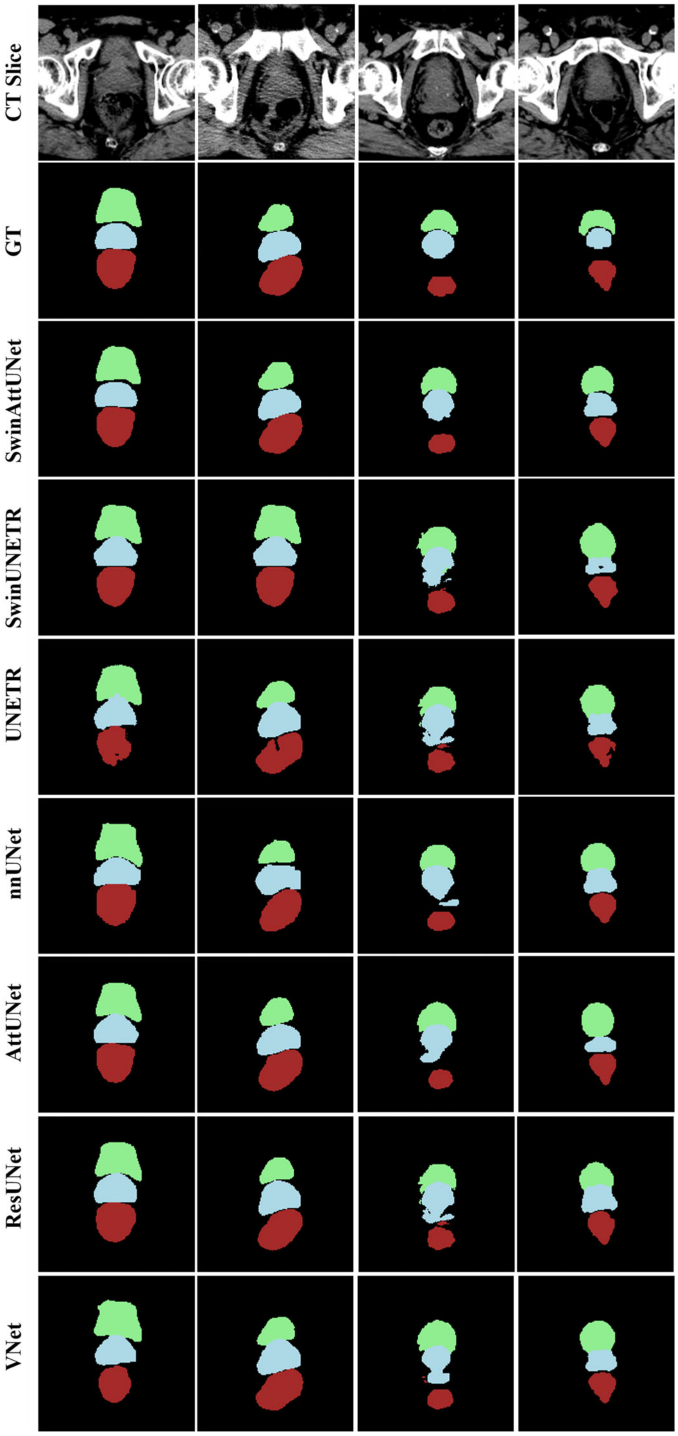

FIGURE 2 Segmentation result from the institutional pelvic dataset. The input CT image of a central slice (Row 1), the ground truth (Row 2), and the predicted segmentation from the SwinAttUNet (Row 3), and all competing networks (Rows 4–9): prostate (blue), bladder (green), and rectum (red).

experiments. The Dice similarity coefficient (DSC) evaluates the overlap of the predicted and ground truth segmentation map

$$DSC = \frac{2\,|P \cap G|}{|P| + |G|} \tag{4}$$

where *P* indicates the predicted segmentation map and *G* denotes the ground truth. A DSC of 1 indicates a perfect segmentation while 0 indicates no overlap at all. Hausdorff distance (HD) measures the largest symmetrical distance between two segmentation maps

$$d_H\,(P, G) = max\{\sup_{p \in P} \inf_{g \in G} d\,(p, g)\,,\, \sup_{g \in G} \inf_{p \in P} d\,(p, g)\} \tag{5}$$

where d(·) represents the Euclidean distance, sup and inf denote supremum and infimum, respectively.

We also include the Average Surface Distance (ASD), Average of all the distances from points on the boundary of the auto-contour to the boundary of the ground truth contour:

$$ASD = \frac{1}{|(S(P)|}\left(\sum_{p\in P} d(S(p), S(G))\right) \qquad (6)$$

where, *d((S(p),S(G))* is the shortest distance of a predicted voxel *S(p)* to the set of ground truth surface voxels, *S(G)*.

2.3.3 | Methods for comparison

The performance of SwinAttUNet was compared against multiple state-of-the-art segmentation models. For FCN-based models, V-Net,[17] ResUNet,[45] AttUNet,[44] and nnUNet[19] were used for comparison using both the institutional and public CT image datasets. For Transformer-based models, UNETR[35] and SwinUNETR[36] were selected for comparison. *P* values were computed using the Mann–Whitney U-test[46] to evaluate statistical significance between contours predicted using SwinAttUNet and the next highest performing network. The significance level was set at 0.05, where $p < 0.05$ indicates statistically significant difference between the two networks.

To validate the effectiveness of the SwinAttUNet architecture, we ran an ablation experiment using the institutional pelvic dataset. We first removed the AG to demonstrate the benefits of AG in the decoding process. We then replaced the parallel encoder with only the CB or TB and compared each of these iterations against the full SwinAttUNet network.

3 | RESULTS

Qualitative comparisons of auto-contours generated with SwinAttUNet, and other networks are presented in Figures 2 and 3. Figure 2 shows results for four example cases based on the institutional pelvic dataset. Contours are shown in the axial view for the ground-truth (GT, Row 2), SwinAttUNet (Row 3), and competing networks (Rows 4–9), for the prostate gland (blue), bladder (green), and rectum (red). Careful inspection of the shapes of these contours and the boundary distances between the different organs (relative to the ground-truth segments, Row 2) shows that the SwinAttUNet performs better than all other networks for all four example cases. Figure 3 shows results for five example cases based on the public dataset (CT-ORG). Contours are depicted for the ground-truth (GT, Row 1), SwinAttUNet (Row 2), and competing networks (Rows 3–8) for the lungs (yellow), liver (green), kidney (cyan), bladder (blue), and pelvic bones (purple). While all networks produce accurate for contours of the liver, lung, kidney, and pelvic bones, the SwinAttUNet is shown to produce the best contours for all organs, including the bladder where discrepancies were noted with the other networks relative to the ground-truth segments.

TABLE 1 Ablation Study: DSC, HD95, and ASD with the different settings for SwinAttUNet on the institutional pelvic dataset.

Organ	Method	DSC (%) ↑	HD95 (mm) ↓	ASD (mm) ↓
Prostate	w/o AG	86.12 ± 1.45	5.23 ± 1.50	1.51 ± 0.63
	w/o CB	85.36 ± 2.43	6.15 ± 1.46	1.62 ± 0.67
	w/o TB	84.69 ± 2.51	5.76 ± 1.43	1.56 ± 0.59
	Full model	**86.54 ± 1.21**	**5.06 ± 1.42**	**1.45 ± 0.57**
Bladder	w/o AG	93.72 ± 4.31	3.18 ± 1.26	0.85 ± 0.61
	w/o CB	93.51 ± 3.32	3.25 ± 1.33	0.93 ± 0.43
	w/o TB	93.24 ± 4.16	3.42 ± 1.37	0.86 ± 0.36
	Full model	**94.15 ± 1.17**	**3.16 ± 0.93**	**0.82 ± 0.12**
Rectum	w/o AG	86.31 ± 2.12	5.74 ± 1.94	1.53 ± 0.51
	w/o CB	86.25 ± 1.83	6.11 ± 2.07	1.61 ± 0.68
	w/o TB	85.49 ± 2.08	5.83 ± 1.85	1.53 ± 0.51
	Full model	**87.15 ± 1.68**	**5.54 ± 1.63**	**1.42 ± 0.38**

Note: Shown are mean ± SD for three experiments for each setting for the prostate gland, bladder, and rectum. The most accurate results are shown in bold font.

Abbreviations: AG, Attention Gate; ASD, Average Surface Distances; CB, CNN Branch; DSC, dice coefficients; HD95, 95% Hausdorff Distance; TB, Transformer Branch.

3.1 | Ablation study

To assess the contribution of the AG, CB, and TB on the segmentation performance, a comparison was performed between the results obtained with the SwinAttUNet (full model) and the network configurations without AG, CB, or TB. Table 1 presents the segmentation results for these three different experiments. The SwinAttUNet (full model) shows superior results for all metrics, DSC, HD95, and ASD for the prostate, bladder, and rectum. The contribution of the various modules of the SwinAttUNet is clearly demonstrated by inferior results when the AG, CB or TB are removed from the network architecture, justifying the need for each module toward the overall accuracy of the SwinAttUNet network.

3.2 | SwinAttUNet network trained on institutional dataset for pelvic organ segmentation

Quantitative results for the SwinAttUNet and other networks trained on the institutional dataset for segmentation of pelvic organs are provided in Table 2. Data are shown for the DSC (%), HD95 (mm) and ASD (mm) for the prostate, bladder, and rectum. *p*-values are also

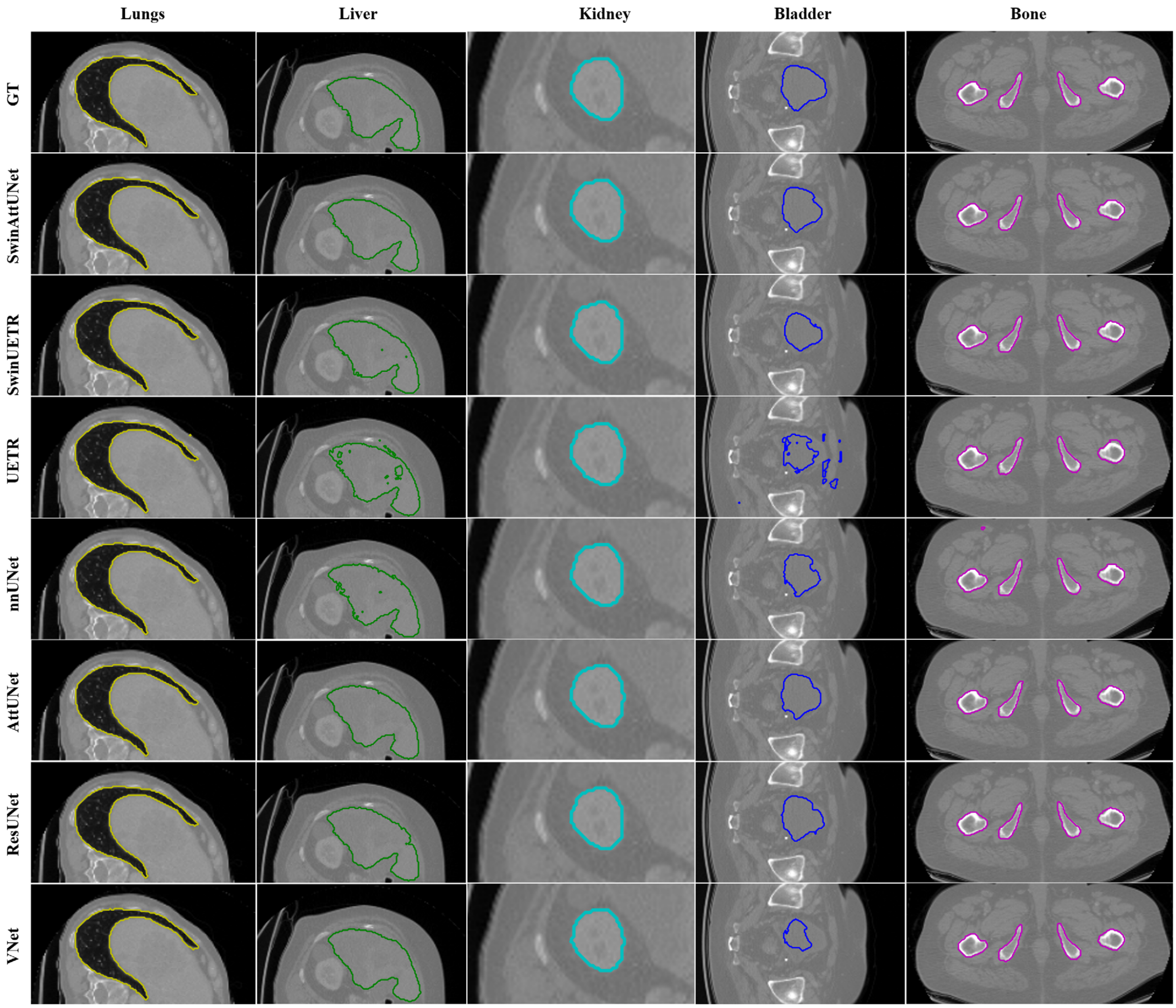

FIGURE 3 Segmentation result from CT-ORG dataset. The selected region of interest of each organ from manual contours (Row 1), the SwinAttUNet (Row 2), and all competing networks (Rows 3–8): lungs (yellow), liver (green), kidney (cyan), bladder (blue), and bones (purple).

included for comparison between the SwinAttUNet and the next highest accuracy network at the 0.05 significance level. For the DSC comparison, the SwinAttUNet outperforms all other networks with values of 86.5% (prostate), 94.2% (bladder) and 87.2% (rectum). The HD95 (mm) values were also lowest for our SwinAttUNet relative to other networks. Statistically significant differences ($p < 0.001$) were observed in the DSC and HD95 (mm) values for our network (SwinAttUNet) versus SwinUNETR for all organs. Apart from the prostate, SwinAttUNet ASD (mm) values outperformed those of all other networks.

For the prostate, the ASD values were 1.40 mm (SwinUNETR) and 1.45 mm (our SwinAttUNet), however, the difference was not statistically significant ($p = 0.076$). Moreover, the standard deviation of the prostate ASD with SwinAttUNet (0.57 mm) was lower than that of SwinUNETR (0.65 mm).

3.3 | SwinAttUNet network trained on public dataset (CT-ORG) for multiple organ segmentation

Quantitative results for the SwinAttUNet and other networks trained on the CT-ORG dataset for segmentation of multiple organs are provided in Table 3. Data are shown for the DSC (%), HD95 (mm) and ASD (mm) for the lungs, liver, kidneys, bladder, and pelvic bones. DSC values are consistently the highest for the SwinAttUNet versus all other networks with values of 97.9% (lungs), 96.2% (liver), 93.7% (kidneys), 88.6% (bladder),

TABLE 2 Quantitative results for networks trained on the institutional pelvic dataset.

Organ	Method	DSC (%) ↑	HD95 (mm) ↓	ASD (mm) ↓
Prostate	V-Net	83.27 ± 2.71	7.75 ± 2.58	2.12 ± 0.73
	ResUNet	84.15 ± 2.61	5.79 ± 1.63	1.74 ± 0.89
	AttUNet	84.26 ± 2.54	5.81 ± 1.56	1.58 ± 0.65
	nnUNet	84.12 ± 2.68	5.83 ± 2.01	1.81 ± 0.79
	UNETR	82.51 ± 4.46	8.92 ± 2.65	2.34 ± 1.01
	SwinUNETR	85.71 ± 2.32	6.10 ± 1.42	**1.40 ± 0.65**
	SwinAttUNet (ours)	**86.54 ± 1.21**	**5.06 ± 1.42**	1.45 ± 0.57
p-values		< 0.001	< 0.001	0.076
Bladder	V-Net	91.56 ± 5.21	6.75 ± 2.01	1.62 ± 0.52
	ResUNet	92.65 ± 4.52	4.46 ± 1.84	1.13 ± 0.24
	AttUNet	93.31 ± 4.23	3.25 ± 1.21	0.87 ± 0.54
	nnUNet	93.46 ± 5.03	4.83 ± 1.59	1.16 ± 0.46
	UNETR	89.37 ± 5.67	6.34 ± 2.56	1.78 ± 0.67
	SwinUNETR	93.62 ± 3.25	3.22 ± 1.14	0.91 ± 0.34
	SwinAttUNet (ours)	**94.15 ± 1.17**	**3.16 ± 0.93**	**0.82 ± 0.12**
p-values		< 0.001	< 0.001	< 0.001
Rectum	V-Net	83.71 ± 3.52	7.12 ± 2.54	2.11 ± 0.61
	ResUNet	86.02 ± 2.34	6.31 ± 2.24	1.58 ± 0.46
	AttUNet	86.63 ± 2.01	5.81 ± 1.95	1.44 ± 0.49
	nnUNet	86.53 ± 2.18	6.14 ± 2.35	1.61 ± 0.62
	UNETR	82.16 ± 4.87	9.76 ± 2.45	2.43 ± 1.10
	SwinUNETR	85.52 ± 2.24	6.12 ± 1.97	1.47 ± 0.61
	SwinAttUNet (ours)	**87.15 ± 1.68**	**5.54 ± 1.63**	**1.42 ± 0.38**
p-values		< 0.001	< 0.001	< 0.001

Note: DSC, HD95, and ASD values represent the mean ± SD for the networks including SwinAttUNet (ours), VNet, ResUNet, AttUNet, nnUNet, UNETR, and SwinUNETR. The most accurate results are shown in bold font. *p*-values are presented to statistically compare the SwinAttUNet against the next highest performing network.

Abbreviations: ASD, Average Surface Distances; DSC, dice coefficients; HD95, 95% Hausdorff Distance.

and 89.3% (pelvic bones). Statistically significant DSC differences ($p < 0.001$) were observed for the SwinAttUNet relative to the SwinUNETR network. Moreover, DSC standard deviations were significantly reduced on segments produced with SwinAttUNet relative to other networks. For instance, the bladder DSC SD was 7.9 mm for SwinAttUNet, while it was >11.5 mm for all other networks. HD95 (mm) values were lowest for our SwinAttUNet relative to other networks for the liver and pelvic bones with statistical significance achieved (against SwinUNETR) for the liver ($p < 0.001$) and pelvic bones ($p = 0.005$). For the lungs, SwinAttUNet HD95 mean value was 5.13 mm while it was slightly better with SwinUNETR (4.95 mm) though the difference was not statistically significant ($p = 0.89$). For the kidneys, SwinAttUNet HD95 mean value was 2.29 mm while it was 2.12 mm for the AttUNet network but not statistically different ($p = 0.65$). For the bladder, the HD95 mean value was 7.68 mm for the V-Net slightly better than 8.23 mm for the SwinAttUNet network for but not statistically different ($p = 0.67$). ASD values were lowest for all organs with our SwinAttUNet network with statistical significance consistently achieved. ASD SD's were also significantly improved with SwinAttUNet. For instance, ASD SDs for liver were reduced to 0.24 mm with SwinAttUNet compared with all other networks where the SDs were generally >0.7 mm, suggesting lower variability and higher consistency in the predicted contours with our network.

4 DISCUSSION

In this work, we propose a U-shaped hierarchically fusing architecture called SwinAttUNet for 3D CT-based multi-organ segmentation. The SwinAttUNet consists of three main components: a convolutional encoder branch for extracting fine local features at different resolutions, a Swin Transformer encoder branch in parallel for enriching global information at each resolution level, and a set

TABLE 3 Quantitative results for networks trained on the public (CT-ORG) multi-organ dataset.

Organ	Method	DSC (%) ↑	HD95 (mm) ↓	ASD (mm) ↓
Liver	V-Net	94.13 ± 2.54	6.48 ± 2.32	1.73 ± 0.63
	ResUNet	94.81 ± 1.81	5.74 ± 3.43	1.42 ± 0.87
	AttUNet	95.23 ± 1.72	4.53 ± 1.67	1.26 ± 0.75
	nnUNet	94.78 ± 1.95	6.21 ± 4.02	1.38 ± 0.84
	UNETR	94.01 ± 2.34	6.83 ± 4.21	5.58 ± 3.54
	SwinUNETR	94.81 ± 2.34	4.85 ± 2.63	1.97 ± 1.65
	SwinAttUNet (ours)	**96.16 ± 0.76**	**2.73 ± 1.19**	**1.08 ± 0.24**
p-values		< 0.001	< 0.001	0.004
Bladder	V-Net	83.24 ± 11.75	**7.68 ± 4.32**	2.56 ± 0.97
	ResUNet	82.48 ± 12.24	9.73 ± 6.85	3.12 ± 1.54
	AttUNet	84.87 ± 11.86	8.56 ± 6.53	2.15 ± 1.17
	nnUNet	85.26 ± 12.58	10.21 ± 8.57	2.53 ± 2.64
	UNETR	82.13 ± 13.65	10.02 ± 5.84	2.86 ± 2.13
	SwinUNETR	83.67 ± 13.15	8.76 ± 6.21	2.24 ± 1.35
	SwinAttUNet (ours)	**88.62 ± 7.91**	8.23 ± 8.01	**1.78 ± 1.21**
p-values		< 0.001	0.673	< 0.001
Lungs	V-Net	95.63 ± 4.36	15.61 ± 8.84	2.89 ± 0.86
	ResUNet	95.82 ± 5.27	6.64 ± 15.76	3.12 ± 4.42
	AttUNet	96.87 ± 5.13	5.99 ± 11.97	3.31 ± 2.37
	nnUNet	95.63 ± 6.85	8.57 ± 5.12	3.53 ± 6.56
	UNETR	93.68 ± 13.64	15.25 ± 19.40	8.37 ± 9.06
	SwinUNETR	95.99 ± 9.30	**4.95 ± 4.37**	2.56 ± 1.63
	SwinAttUNet (ours)	**97.90 ± 0.80**	5.13 ± 4.11	**1.88 ± 1.45**
p-values		< 0.001	0.893	< 0.001
Kidney	V-Net	88.15 ± 3.25	4.45 ± 1.87	2.14 ± 3.32
	ResUNet	92.03 ± 3.40	3.26 ± 1.32	0.95 ± 0.75
	AttUNet	92.85 ± 3.73	**2.12 ± 1.14**	0.83 ± 0.58
	nnUNet	93.41 ± 3.92	3.11 ± 5.73	1.01 ± 0.78
	UNETR	88.96 ± 4.88	6.36 ± 9.03	3.18 ± 3.59
	SwinUNETR	91.87 ± 3.41	3.37 ± 1.24	0.94 ± 0.66
	SwinAttUNet (ours)	**93.74 ± 2.25**	2.29 ± 1.47	**0.71 ± 0.43**
p-values		< 0.001	0.653	< 0.001
Bones	V-Net	86.45 ± 2.17	8.76 ± 3.21	2.22 ± 3.28
	ResUNet	88.61 ± 4.95	5.58 ± 5.87	1.44 ± 1.23
	nnUNet	88.63 ± 4.57	5.67 ± 6.12	2.19 ± 2.67
	UNETR	86.85 ± 6.39	8.78 ± 9.03	4.72 ± 4.90
	SwinUNETR	88.97 ± 4.80	5.63 ± 6.03	2.43 ± 2.44
	SwinAttUNet (ours)	**89.31 ± 3.87**	**5.31 ± 1.25**	**1.21 ± 1.11**
p-values		< 0.001	0.005	0.023

Note: The Table shows results for the DSC, HD95, and ASD (mean ± SD) for the networks, including SwinAttUNet (ours), VNet, ResUNet, AttUNet, nnUNet, UNETR, and SwinUNETR. The most accurate results are shown in bold font. *p*-values are presented to statistically compare the SwinAttUNet against the next highest performing network.

Abbreviations: ASD, Average Surface Distances; DSC, dice coefficients; HD95, 95% Hausdorff Distance.

of AG-regulated, up-sampling convolutional blocks for reconstruction of features into an *N*-class segmentation. Our network is novel in that the transformer layers effectively capture global information in parallel with the CNN layers for each resolution level, overcoming the receptive field limitations of pure fully convolutional networks (FCN's). Additionally, the novel AGs enable effective interaction of extracted features from different resolution levels, as evidenced by the ablation study. The proposed network demonstrates promising segmentation performance compared to current state of-the-art methods for auto-segmentation of organ contours in multiple regions of the body including the pelvis, thorax, and gastro-intestinal (GI). The superior DSC, HD95, and ASD results of our proposed network highlight the advantages of parallelizing the CNN and Swin Transformer layers in the encoding stage for CT-based multi-organ segmentation.

Relative to our network other CNN-based models with more complex architectures and ground truth segmentations based on multi-modal imaging information were able to achieve similar accuracies. An example of this is a study by Dong et al.,[47] where the investigators utilized a Cycle-GAN for 3D CT-to-synthetic MRI synthesis and trained the segmentation network on the synthetic MRI (sMRI) scans. They reported Dice scores of 0.87 ± 0.04 for the prostate, and 0.95 ± 0.03 for the bladder using 140 pelvic image datasets. Similarly, other investigators used networks such as GAN for CT-to-sMRI synthesis[48] (with Dice scores of 0.90 ± 0.05 for rectum) and 2D organ localization networks[27] (with Dice scores of 0.95 ± 0.02 for bladder). In the context of postoperative prostate cancer, Balagopal et al.[49] developed a deep learning network (2D U-Net) for auto-segmentation of the clinical target volume (CTV) incorporating uncertainty. The training dataset consisted of 340 patients with post-operative prostate cancer, with ground-truth contours drawn by physicians. A DSC value of 0.87 was reported for a holdout dataset (50 patient CT images). Balagopal et al.[50] also developed a deep learning network (based on a 3D-CNN), PSA-Net, for segmentation of the CTV trained to incorporate differences in physician preferences during segmentation. For training, 373 postoperative prostate cancer CT image datasets were employed. Questions such as consistency in physician contouring preferences and whether inter-user variation in segmentation affects treatment outcomes were addressed. DSC values of 0.87 were reported for their network.

There are a few limitations to be noted. We trained our network on two independent (institutional and public datasets) training datasets because the ground-truth labels/contours were not available for the same organs on these datasets. If contours were available for the same set of organs, it would have become feasible to train the network with just one training dataset, which may be more practical for clinical application. The generalization error of a network is best tested using "unseen" datasets from an independent institution, as it tests the robustness of the network to variation associated with multiple factors, such as image intensity and contrast, patient anatomy, inter-observer differences in ground truth contours of expert annotators, etc.

As part of future research, we intend to evaluate the network using 'unseen' datasets from independent institutions. We will also incorporate advanced techniques/networks to enhance segmentation accuracy of SwinAttUNet. For instance, we propose to extend the parallel CNN and Transformer into the decoding process, which has potential to increase segmentation accuracy. Our network is efficient. Apart from the training phase (which requires >10 h but is done prior to clinical application), the network is fast, requiring about 5 s/case for routine multi-organ contour generation, thereby facilitating auto-segmentation for procedures such as on-table adaptive treatment. We are investigating techniques to automatically detect and correct outliers from either manually (user-defined) or automatically generated contours using this network. These tools are likely to be of value toward the overall quality assurance of target and normal organ segmentation in radiation treatment planning.

5 | CONCLUSION

This study introduces a segmentation network that leverages a novel parallel encoding approach, combining advantages from both CNN and Transformer encoders with self-attention, for multi-organ, 3D-CT auto-segmentation. The proposed method generally surpasses VNet and other state-of-the-art models for auto-contouring. The network exhibits potential as an accurate and efficient tool for facilitating automatic segmentation for procedures, such as radiation treatment planning.

ACKNOWLEDGEMENTS

This work was supported in part by a grant from Varian Medical Systems (Siemens Healthineers, Palo Alto, CA, USA)

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

REFERENCES

1. Bi WL, Hosny A, Schabath MB, et al. Artificial intelligence in cancer imaging: clinical challenges and applications. *CA Cancer J Clin*. 2019;69(2):127-157.
2. Thor M, Olsson C, Deasy J, et al. Dose-response relationships for four gastrointestinal symptom groups in prostate cancer radiation therapy. *Int J Radiat Oncol Biol Phys*. 2015;93(3):S52.
3. Zeleznik R, Weiss J, Taron J, et al. Deep-learning system to improve the quality and efficiency of volumetric heart segmentation for breast cancer. *NPJ Digital Med*. 2021;4(1):43.

4. Hadjiiski L, Cha K, Chan HP, et al. AAPM task group report 273: recommendations on best practices for AI and machine learning for computer-aided diagnosis in medical imaging. *Med Phys*. 2023;50(2):e1-e24.
5. El Naqa I, Das S, The role of machine and deep learning in modern medical physics. 2020.
6. Feng M, Valdes G, Dixit N, Solberg TD. Machine learning in radiation oncology: opportunities, requirements, and needs. *Front Oncol*. 2018;8:110.
7. Li X, Bagher-Ebadian H, Gardner S, et al. An uncertainty-aware deep learning architecture with outlier mitigation for prostate gland segmentation in radiotherapy treatment planning. *Med Phys*. 2023;50(1):311-322.
8. van Herk M. Errors and margins in radiotherapy. *Semin Radiat Oncol*. 2004;14(1):52-64.
9. Gudmundsson E, Straus CM, Li F. Deep learning-based segmentation of malignant pleural mesothelioma tumor on computed tomography scans: application to scans demonstrating pleural effusion. *J Med Imaging*. 2020;7(1):012705-012705.
10. Liang X, Bibault JE, Leroy T, et al. Automated contour propagation of the prostate from pCT to CBCT images via deep unsupervised learning. *Med Phys*. 2021;48(4):1764-1770.
11. Breto AL, Spieler B, Zavala-Romero O, et al. Deep learning for per-fraction automatic segmentation of gross tumor volume (GTV) and organs at risk (OARs) in adaptive radiotherapy of cervical cancer. *Front Oncol*. 2022;12:854349.
12. Interian Y, Rideout V, Kearney VP, et al. Deep nets vs expert designed features in medical physics: an IMRT QA case study. *Med Phys*. 2018;45(6):2672-2680.
13. Glocker B, Pauly O, Konukoglu E, Criminisi A. Joint classification-regression forests for spatially structured multi-object segmentation. Paper presented at: Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7–13, 2012, Proceedings, Part IV 122012.
14. Chen C, Zheng G. Fully automatic segmentation of AP pelvis X-rays via random forest regression with efficient feature selection and hierarchical sparse shape composition. *Comput Vis Image Underst*. 2014;126:1-10.
15. Gao Y, Shao Y, Lian J, Wang AZ, Chen RC, Shen D. Accurate segmentation of CT male pelvic organs via regression-based deformable models and multi-task random forests. *IEEE Trans Med Imaging*. 2016;35(6):1532-1543.
16. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. Paper presented at: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 182015.
17. Milletari F, Navab N, Ahmadi S-A. V-net: fully convolutional neural networks for volumetric medical image segmentation. Paper presented at: 2016 fourth international conference on 3D vision (3DV), Stanford, CA, USA, 2016:565-571.
18. Kazemifar S, Balagopal A, Nguyen D, et al. Segmentation of the prostate and organs at risk in male pelvic CT images using deep learning. *Biomed Phys Eng Express*. 2018;4(5):055003.
19. Isensee F, Jaeger PF, Kohl SA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods*. 2021;18(2):203-211.
20. Cardenas CE, Yang J, Anderson BM, Court LE, Brock KB. Advances in auto-segmentation. *Semin Radiat Oncol*., 2019;29(3):185-197.
21. Cui S, Tseng HH, Pakela J. Ten Haken RK, El Naqa I. Introduction to machine and deep learning for medical physicists. *Med Phys*. 2020;47(5):e127-e147.
22. Seo H, Huang C, Bassenne M, Xiao R, Xing L. Modified U-Net (mU-Net) with incorporation of object-dependent high level features for improved liver and liver-tumor segmentation in CT images. *IEEE Trans Med Imaging*. 2019;39(5):1316-1325.
23. Yu C, Anakwenze CP, Zhao Y, et al. Multi-organ segmentation of abdominal structures from non-contrast and contrast enhanced CT images. *Sci Rep*. 2022;12(1):19093.
24. Zeleznik R, Foldyna B, Eslami P, et al. Deep convolutional neural networks to predict cardiovascular risk from computed tomography. *Nat Commun*. 2021;12(1):715.
25. Zavala-Romero O, Breto AL, Xu IR, et al. Segmentation of prostate and prostate zones using deep learning: a multi-MRI vendor analysis. *Strahlenther Onkol*. 2020;196:932-942.
26. Schipaanboord B, Boukerroui D, Peressutti D, et al. An evaluation of atlas selection methods for atlas-based automatic segmentation in radiotherapy treatment planning. *IEEE Trans Med Imaging*. 2019;38(11):2654-2664.
27. Balagopal A, Kazemifar S, Nguyen D, et al. Fully automated organ segmentation in male pelvic CT images. *Phys Med Biol*. 2018;63(24):245015.
28. Pan S, Lei Y, Wang T, et al. Male pelvic multi-organ segmentation using token-based transformer Vnet. *Phys Med Biol*. 2022;67(20):205012.
29. Kearney V, Chan JW, Wang T, Perry A, Yom SS, Solberg TD. Attention-enabled 3D boosted convolutional neural networks for semantic CT segmentation using deep supervision. *Phys Med Biol*. 2019;64(13):135001.
30. Wang S, Liu M, Lian J, Shen D. Boundary coding representation for organ segmentation in prostate cancer radiotherapy. *IEEE Trans Med Imaging*. 2020;40(1):310-320.
31. Luo W, Li Y, Urtasun R, Zemel R. Understanding the effective receptive field in deep convolutional neural networks. *Adv Neural Inf Process Syst*. 2016;29.
32. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16×16 words: Transformers for image recognition at scale. arXiv preprint arXiv:201011929. 2020.
33. Chen J, Lu Y, Yu Q, et al. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:210204306*. 2021.
34. Cao H, Wang Y, Chen J, et al. Swin-unet: unet-like pure transformer for medical image segmentation. Paper presented at: Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part III2023.
35. Hatamizadeh A, Tang Y, Nath V, et al. Unetr: transformers for 3d medical image segmentation. Paper presented at: Proceedings of the IEEE/CVF winter conference on applications of computer vision, 2022.
36. Tang Y, Yang D, Li W, et al. Self-supervised pre-training of swin transformers for 3d medical image analysis. Paper presented at: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022.
37. Zhang Y, Liu H, Hu Q. Transfuse: fusing transformers and cnns for medical image segmentation. Paper presented at: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 242021.
38. Yoo J, Kim T, Lee S, Kim SH, Lee H, Kim TH. Enriched CNN-Transformer Feature Aggregation Networks for Super-Resolution. Paper presented at: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023.
39. He K, Gan C, Li Z, et al. Transformers in medical image analysis: a review. *Intelligent Medicine*. 2022;23:59-78.
40. Liu C, Gardner SJ, Wen N, et al. Automatic segmentation of the prostate on CT images using deep neural networks (DNN). *Int J Radiat Oncol Biol Phys*. 2019;104(4):924-932.
41. Rister B, Yi D, Shivakumar K, Nobashi T, Rubin DL. CT-ORG, a new dataset for multiple organ segmentation in computed tomography. *Sci Data*. 2020;7(1):381.
42. Liu Z, Lin Y, Cao Y, et al. Swin transformer: hierarchical vision transformer using shifted windows. Paper presented at: Proceedings of the IEEE/CVF international conference on computer vision, 2021.

43. He K, Zhang X, Ren S, Sun J. Identity mappings in deep residual networks. Paper presented at: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 1420 16.
44. Oktay O, Schlemper J, Folgoc LL, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:180403999*. 2018.
45. Gao Y, Zhou M, Liu D, Yan Z, Zhang S, Metaxas DN. A data-scalable transformer for medical image segmentation: architecture, model efficiency, and benchmark. *arXiv preprint arXiv:220300131*. 2022.
46. McKnight PE, Najab J. Mann-Whitney U test. *The Corsini encyclopedia of psychology*. 2010:1-1.
47. Dong X, Lei Y, Tian S, et al. Synthetic MRI-aided multi-organ segmentation on male pelvic CT using cycle consistent deep attention network. *Radiother Oncol*. 2019;141:192-199.
48. Lei Y, Wang T, Tian S, et al. Male pelvic multi-organ segmentation aided by CBCT-based synthetic MRI. *Phys Med Biol*. 2020;65(3):035013.
49. Balagopal A, Nguyen D, Morgan H, et al. A deep learning-based framework for segmenting invisible clinical target volumes with estimated uncertainties for post-operative prostate cancer radiotherapy. *Med Image Anal*. 2021;72:102101.
50. Balagopal A, Morgan H, Dohopolski M, et al. PSA-Net: deep learning–based physician style–aware segmentation network for postoperative prostate cancer clinical target volumes. *Artif Intell Med*. 2021;121:102195.

How to cite this article: Li C, Bagher-Ebadian H, Sultan RI, et al. A new architecture combining convolutional and transformer-based networks for automatic 3D multi-organ segmentation on CT images. *Med Phys*. 2023;50:6990–7002. https://doi.org/10.1002/mp.16750