Dual Stream Fusion U-Net Transformers for 3D Medical Image Segmentation

Seungkyun Hong*, Sunghyun Ahn*, Youngwan Jo, Sanghyun Park†
Department of Computer Science, Yonsei University
Seoul, Republic of Korea
{highsk, skd, jyy1551, sanghyun}@yonsei.ac.kr

***Abstract*—Medical image segmentation is a crucial ongoing issue in clinical applications for differentiating lesions and segmenting various organs to extract relevant features. Many recent studies have combined transformers, which enable global context modeling leveraging self-attention, with U-Nets to distinguish organs in complex volumetric medical images such as 3-dimensional (3D) Computed Tomography (CT) and Magnetic Resonance Imaging (MRI) images. In this study, we propose a Dual Stream fusion U-NEt TRansformers (DS-UNETR) comprising a Dual Stream Attention Encoder (DS-AE) and Bidirectional All Scale Fusion (Bi-ASF) module. We designed the DS-AE that extracts both spatial and channel features in parallel streams to better understand the relation between channels. When transferring the extracted features from the DS-AE to the decoder, we used the Bi-ASF module to fuse all scale features. We achieved an average Dice similarity coefficient (Dice score) improvement of 0.97% and a 95% Hausdorff distance (HD95), indicating an improvement of 7.43% compared to that for a state-of-the-art model on the Synapse dataset. We also demonstrated the efficiency of our model by reducing the space and time complexity with a decrease of 80.73% in parameters and 78.86% in FLoating point OPerationS (FLOPS). Our proposed model, DS-UNETR, shows superior performance and efficiency in terms of segmentation accuracy and model complexity (both space and time) compared to existing state-of-the-art models on the 3D medical image segmentation benchmark dataset. The approach of our proposed model can be effectively applied in various medical big data analysis applications.**

***Index Terms*—Channel attention, Multi-scale fusion, Swin Transformer, 3D medical image segmentation**

I. Introduction

Medical image segmentation is a crucial task for analyzing anatomical structures in the body to facilitate computer-aided diagnosis, image-guided surgery, and other medical procedures. Particularly, 3-dimensional (3D) volumetric segmentation is a fundamental problem in medical imaging for numerous applications including the identification of tumors and the determination of the location of organs for diagnostic purposes [1]–[3]. One of the best-known structures for medical image segmentation using deep learning models is U-Net [4]. The convolutional neural network (CNN)-based approach adopted in U-Net has achieved significant success in several vision tasks such as classification and object detection, and has also shown ‘ results in medical image segmentation. However, despite the excellent representational power of CNN-based methods, each convolution kernel can only focus on a sub-region of the entire image due to inherent inductive bias, resulting in a loss of global context and inability to build long-range dependencies [5]–[7].

Recently, the transformer [8] attention mechanism used in natural language processing [9] has been applied to vision tasks through the Vision Transformer (ViT) [10], overcoming the limitations of CNNs. ViT has gained popularity recently due to its ability to encode long-range dependencies, leading to promising results in various vision tasks. However, it has a limitation of quadratic complexity in self-attention with respect to the dimensions of input images. To alleviate this problem, Liu et al. proposed the Swin transformer [11], which uses shifted windows for attention and constructs hierarchical feature maps. This structure admits linear complexity with respect to image size, thus overcoming the limitations of ViT. Transformer-based models did not initially attract enough attention in medical image segmentation, but recent studies, such as those on TransUNet [12] and TransFuse [13], have demonstrated the significant potential of transformers in medical image segmentation by utilizing them in encoders [7].

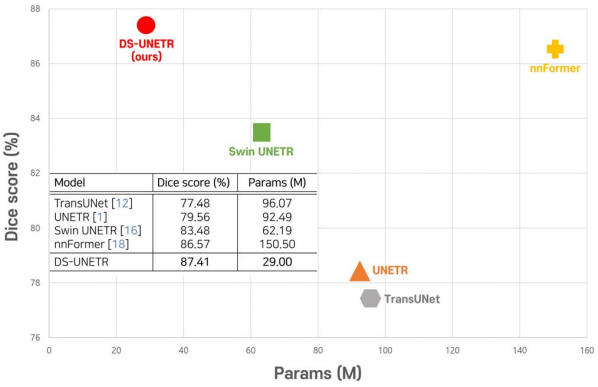
Cao et al. applied the first pure transformer approach, Swin-unet [14], in 2-dimensional (2D) medical image segmentation by using Swin transformer blocks in all layers of the encoder and decoder [15]. Hatamizadeh et al. and Zhou et al. proposed a hybrid approach called Swin UNETR [16] and nnFormer [17], respectively. Hatamizadeh et al. used the Swin transformer as the encoder and CNN-based decoder for 3D medical image segmentation, and achieved the best performance in the BraTS 2021 [18] segmentation challenge. Zhou et al. used an attention mechanism in both the encoder and decoder, and replaced traditional concatenation/summation with attention during skip connections. Additionally, they employed convolution in the embedding, down-sampling, and up-sampling processes, adopting a different approach from that in Swin UNETR. However, most U-shaped 3D medical image segmentation models, including nnFormer, have primarily focused on spatial information, although the channel count of feature maps doubles in the encoder structure. Convolutional encoders simply combine channel information during the convolution process, and transformer encoders exclusively perform spatial attention mechanisms, potentially overlooking crucial information along the channel axis. These issues can lead to the underutilization of contextual information in images. Furthermore, because

* Equal contribution
† Corresponding author: Sanghyun Park (sanghyun@yonsei.ac.kr)

2375-9356/24/$31.00 ©2024 IEEE
DOI 10.1109/BigComp60711.2024.00054

skip connections deliver features from the same layer of the encoder to the corresponding layer of the decoder, each decoder layer lacks the utilization of fused information from various features extracted in the encoder layer.

To address these issues, we propose the Dual Stream fusion U-NEt TRansformers (DS-UNETR). Unlike previous studies [7], [14], [16], we parallelized the spatial attention stream and channel attention stream in the encoder of the U-Net, and fused spatial and channel information at each stage, capturing not only the spatial features, but also the inter-channel relations. Additionally, we improved the representation power of 3D images by fusing all scales of features from the encoder leveraging a Bidirectional All Scale Fusion (Bi-ASF) module. We also minimized the increase in space complexity of the model by using the Window Multi-Head Self-Attention (W-MSA) and Shifted Window Multi-Head Self-Attention (SW-MSA) of the Swin transformer during spatial attention, thereby reducing the number of parameters. Our goal is to achieve improved performance in the segmentation of 3D medical images through this approach.



Model	Dice score (%)	Params (M)
TransUNet [12]	77.48	96.07
UNETR [1]	79.56	92.49
Swin UNETR [16]	83.48	62.19
nnFormer [18]	86.57	150.50
DS-UNETR	87.41	29.00

Fig. 1: Dice similarity coefficient (Dice score) and parameters comparison on the Synapse dataset.

Our contributions can be summarized as follows:

- We propose the Dual Stream Attention Encoder (DS-AE) that combines the spatial attention stream and channel attention stream of the encoder in U-Net in parallel, and fuses spatial and channel information at each stage.
- We propose an effective Bi-ASF module that fuses both coarse-grained and fine-grained information in a bidirectional manner extracted at each stage of the DS-AE.
- Our proposed model, DS-UNETR, achieved state-of-the-art performance in terms of the average Dice similarity coefficient (Dice score) on the Synapse multi-organ segmentation dataset, while also significantly reducing the model's space and time complexity in terms of parameters and FLoating point OPerationS (FLOPS) when compared to the latest state-of-the-art model, demonstrating the efficiency of the model's structure, as shown in Fig. 1.

II. Proposed Method

A. The overview of the DS-UNETR framework

Fig. 2 represents the structure of our proposed DS-UNETR. It comprises the DS-AE (comprising spatial attention stream, channel attention stream, and a module that fuses spatial and channel information at each stage), the Bi-ASF module for feature fusion at all scales, and a decoder that uses Swin transformer blocks (Swin block) for each stage. We constructed the DS-AE, utilizing Swin blocks for performing spatial attention in one stream and Channel Multi-head Self-Attention (C-MSA) blocks for performing channel attention in the other stream in parallel. At each stage in the DS-AE, a Fusion block is used to fuse the attention features extracted from these two blocks, as shown in Fig. 3. This allows us to better understand not only the spatial features of 3D medical images, but also the inter-channel relations. A detailed explanation of the DS-AE is provided in section Section II-B. Additionally, we designed the Bi-ASF module that fuses features at all scales in both the bottom-up and top-down directions, allowing local and global contexts of fused features at each stage of the DS-AE to be effectively transferred to the decoder. A detailed explanation of the Bi-ASF module is provided in section Section II-C.

The input $z \in \mathbb{R}^{H \times W \times D \times C}$ of DS-UNETR is partitioned into patches of size *(h, w, d)* and transferred through a linear embedding layer, resulting in $z^{in} \in \mathbb{R}^{\frac{H}{h} \times \frac{W}{w} \times \frac{D}{d} \times 32}$, which is then fed into the spatial attention stream and channel attention stream. Here, z^{in} is an input of the DS-AE. Additionally, patch merging and expanding are performed to reduce and increase the resolution of the input by a factor of 2. ResBlock in Fig. 2 comprises two blocks of convolution with a kernel size of (3,3,3) and stride of 1, batch normalization, and Leaky ReLU. It has a structure that performs Leaky ReLU again after the input feature and residual connection. At the final layer of the model, the segmentation result is outputted through a 3D convolution layer with a $(1 \times 1 \times 1)$ kernel size in the segmentation head.

B. DS-AE

The DS-AE comprises two parallel streams of spatial and channel attention, and a fusion module that fuses spatial and channel information at each stage, as shown in Fig. 2.

In the DS-AE, at each stage of the spatial attention stream and channel attention stream, the Swin block and C-MSA block are performed three times each in parallel. The Swin block and C-MSA block are illustrated in Fig. 3.

In the first stage of each stream, the input $z^{in} \in \mathbb{R}^{H' \times W' \times D' \times C'}$ is reshaped into $s^0, c^0 \in \mathbb{R}^{H'W'D' \times C'}$, which are used as inputs to the Swin block and C-MSA block, respectively. H', W', D' represent $\frac{H}{h}, \frac{W}{w}, \frac{D}{d}$, respectively. Here, *h, w, and d* double in size through patch merging and halve in size through patch expanding. C' represents the number of channels in each stage. The initial value of C' is 32 in the DS-AE, and it also doubles or halves in size through patch merging and patch expanding. The l^{th} spatial and channel attention are

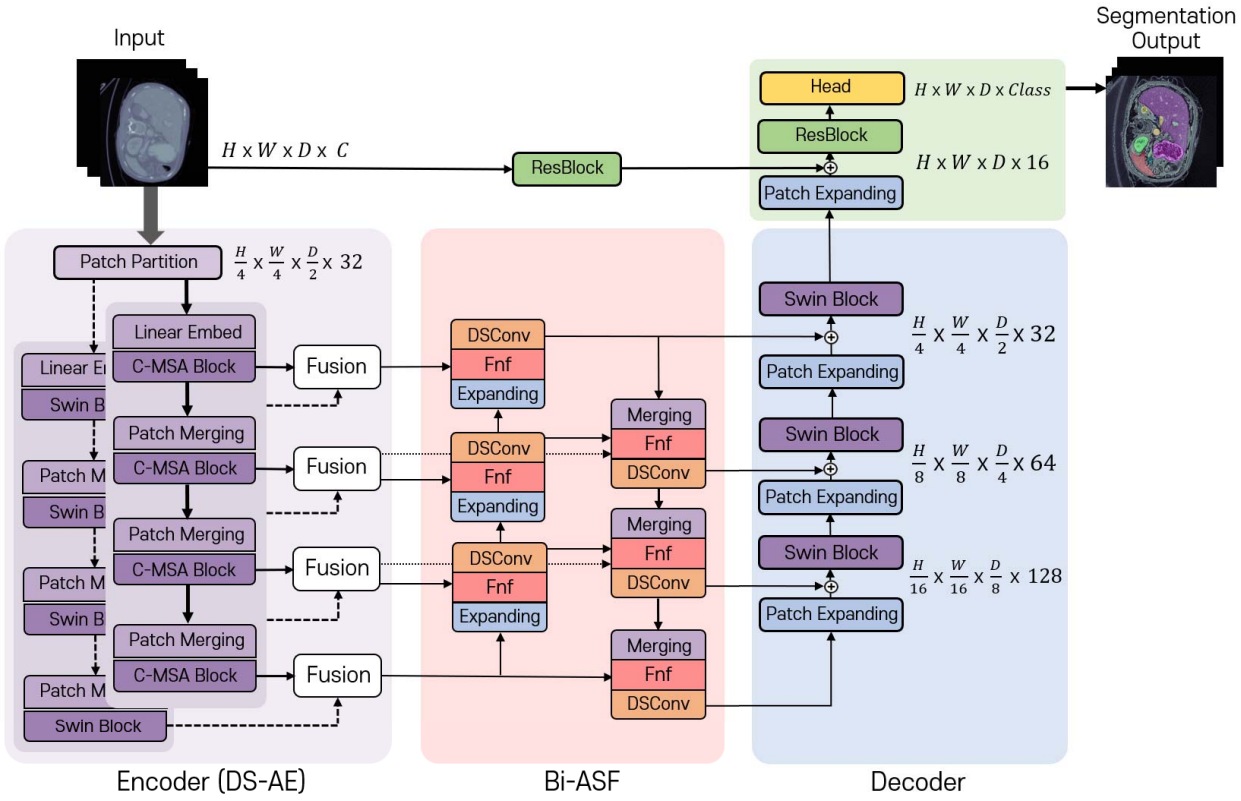


Fig. 2: The overview of the DS-UNETR framework. In DS-AE, the outputs of the Swin block and C-MSA block in each stage of each stream are fused by the Fusion block. In Bi-ASF, the Fnf is performed on the features received from DS-AE, followed by depth-wise separable convolution (DSConv).

calculated as in (1) and (2), and l represents the order of each attention block, from 1 to 12.

$$s_1^l = \text{W-MSA}(\text{LN}(s^l)) + s^l, \quad s_2^l = \text{MLP}(\text{LN}(s_1^l)) + s_1^l,$$
$$s_3^l = \text{SW-MSA}(\text{LN}(s_2^l)) + s_2^l, \; s^{l+1} = \text{MLP}(\text{LN}(s_3^l)) + s_3^l, \tag{1}$$

where s^{l+1} represents the outputs of the Swin block.

$$c_1^l = \text{C-MSA}(\text{LN}(c^l)) + c^l, \;\; c^{l+1} = \text{MLP}(\text{LN}(c_1^l)) + c_1^l, \tag{2}$$

where c^{l+1} represents the outputs of the C-MSA block.

Channel attention computes the attention over channels by multiplying the value with the channel attention map generated by transposing the query and multiplying with the key. The channel attention is computed as shown in (3).

$$\text{Channel Attention}(Q, K, V) = V \cdot \text{Softmax}(\frac{Q^T K}{\sqrt{d}}), \tag{3}$$

where Q, K, and V represent queries, keys, and values, respectively, and d represents the dimension of the query and key.

C-MSA performs h times channel self attention instead of performing a single attention over the entire dimension (C') of keys, queries, and values, and then concatenates the results. C-MSA is computed as shown in (4).

$$\text{C-MSA}(Q, K, V) = \text{Concat}(\text{head}_1, ..., \text{head}_h)W^O, \tag{4}$$

where $\text{head}_i = \text{ChannelAttention}(QW_i^Q, KW_i^K, VW_i^V)$, $W_i^Q, W_i^K, W_i^V, W^O \in \mathbb{R}^{hd \times d_{model}}$ are parameter matrices of the projections, and d_{model} is equal to C'.

In this work, we employ h = 4 parallel attention layers, or heads. For each of these, we use the formula $d = d_{model}/h$, where h is the number of heads in the multi-head attention mechanism. When the stage is 1, 2, 3, or 4, the value of d_{model} (which is equivalent to C') becomes 32, 64, 128, or 256, respectively, and as a result, the value of d is set to 8, 16, 32, or 64, respectively.

The third output of each stage, s^i and c^i $(i = 3, 6, 9, 12)$, are processed through the Fusion block of DS-AE to generate $z^i \in \mathbb{R}^{H' \times W' \times D' \times C'}$, as shown in (5). The output z^i of the Fusion block is used as an input to the Bi-ASF module.

$$z^i = \text{ResBlock}(\text{Reshape}(\text{Concat}(\text{CR}(s^i), \text{CR}(c^i)))). \tag{5}$$

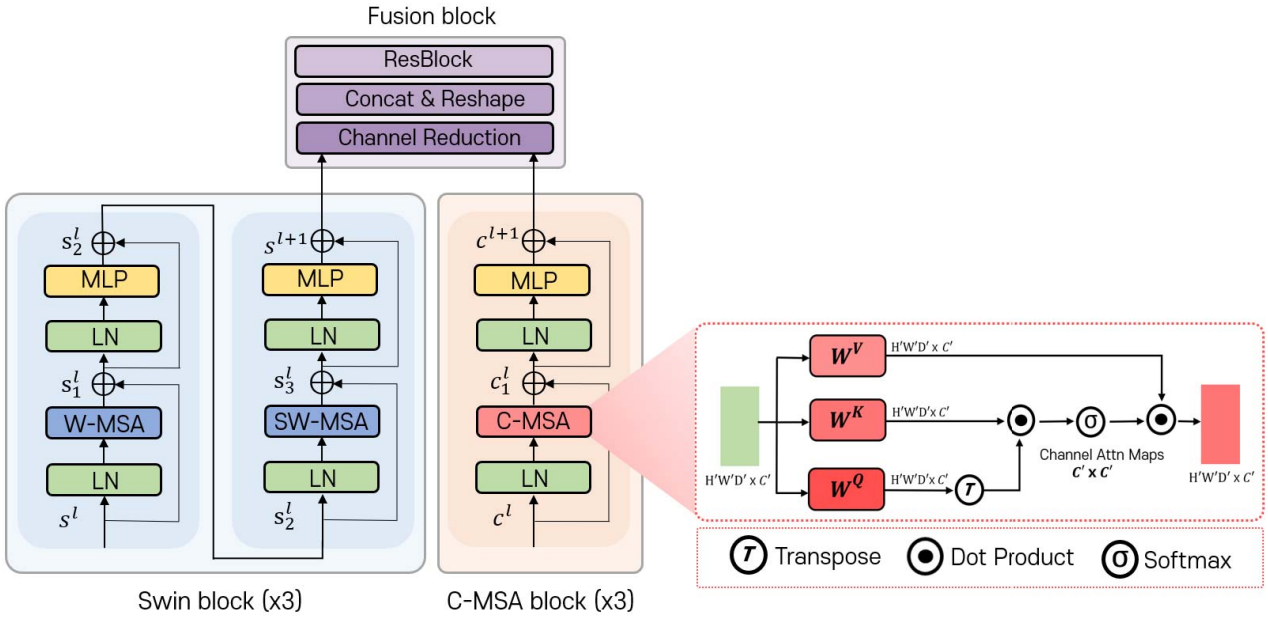


Fig. 3: The Swin block, C-MSA block, and Fusion block of each stage in the encoder. The Swin block and C-MSA block are repeated three times in each stage, and the outputs of each block are used as inputs to the Fusion block. The Fusion block includes several steps such as channel reduction, concatenation, reshaping, and ResBlock operation to effectively fuse the outputs. Abbreviations are: LN: Layer Normalization, MLP: Multi-Layer Perceptron.

The Fusion block comprises linear layers for channel reduction (CR), concatenation, reshape, and ResBlock. After channel reduction, s^i and c^i become $H'W'D' \times C'/2$, and after concatenation, they become $H'W'D' \times C'$. Then, after the reshape process, they become $H' \times W' \times D' \times C'$, which is used as the input to the ResBlock. ResBlock comprises two blocks of convolution with a kernel size of (3,3,3) and stride of 1, batch normalization, and Leaky ReLU. It has a structure that executes the Leaky ReLU again after the residual connection. The input channel size and output channel size are the same.

C. Bi-ASF module

Inspired by [19], we designed an effective Bi-ASF module that fuses all scale features extracted in a bidirectional manner to efficiently transfer local context extracted from the higher stage of the encoder and global context extracted from the lower stage of the DS-AE in the decoder, as shown in Fig. 2.

We execute the Bi-ASF module three times in our model. In each stage of the Bi-ASF module, we perform the Fnf [19] on the features extracted from each stage of the encoder using (6). The result of the Fnf is used as the input of the depth-wise separable convolution [20].

The Fnf is a simple and efficient approach to combine feature maps by normalizing the values of feature maps at various scales and then summing them. In order to perform all scale fusion, we performed the Fnf in the bottom-up and top-down directions.

$$\text{Fnf}\ (I) = \sum_i \frac{w_i}{\epsilon + \sum_j w_j} \cdot I_i, \tag{6}$$

where w denotes a learnable weight, I denotes an input feature map, i and j denote the number of input features, $w_i \geq 0$ is ensured by applying a ReLU after each w_i, and ϵ = 0.0001 is set to avoid numerical instability [19].

D. Loss function

We use a sum of the commonly used soft dice loss [22] and cross-entropy loss as the loss function. It is defined as follows:

$$\mathcal{L}(G,P) = \left(1 - \frac{2}{C}\sum_{c=1}^{C}\frac{\sum_{v=1}^{V} G_{v,c} \cdot P_{v,c}}{\sum_{v=1}^{V} G_{v,c}^2 + \sum_{v=1}^{V} P_{v,c}^2}\right) - \sum_{c=1}^{C}\sum_{v=1}^{V} G_{v,c} \log P_{v,c}, \tag{7}$$

where C denotes the total number of classes, c denotes the index of classes, V denotes the total number of voxels, and v denotes the index of voxels. $G_{v,c}$ and $P_{v,c}$ represent the ground truths and output probabilities at the v^{th} voxel for the c^{th} class, respectively.

III. Experiments

A. Datasets

We conducted experiments on two datasets to compare our method with previous methods. Additionally, to ensure

TABLE I: Comparison on the abdominal multi-organ segmentation (Synapse) dataset. Abbreviations are: Spl: spleen, RKid: right kidney, LKid: left kidney, Gal: gallbladder, Liv: liver, Sto: stomach, Aor: aorta, Pan: pancreas. Best results are **bolded**. Best seconds are underlined.

Methods	Spl	RKid	LKid	Gal	Liv	Sto	Aor	Pan	Average		Params (M)	FLOPS (G)
									Dice score ↑	HD95 ↓		
TransUNet [12]	85.08	77.02	81.87	63.16	94.08	75.62	87.23	55.86	77.48	31.69	96.07	88.91
Swin-Unet [14]	90.66	79.61	83.28	66.53	94.29	76.60	85.47	56.58	79.13	21.55	-	-
UNETR [1]	87.81	84.80	85.66	60.56	94.46	73.99	89.99	59.25	79.56	22.97	92.49	75.76
MISSFormer [21]	91.92	82.00	85.21	68.65	94.41	80.81	86.99	65.67	81.96	18.20	-	-
Swin UNETR [16]	95.37	86.26	86.99	66.54	95.72	77.01	91.12	68.80	83.48	10.55	62.19	350.60
nnFormer [17]	90.51	86.25	86.57	70.17	**96.84**	**86.83**	**92.04**	**83.35**	86.57	10.63	150.50	213.40
DS-UNETR	**95.81**	**87.52**	**87.31**	**74.65**	96.38	86.04	91.55	80.02	**87.41**	**9.84**	**29.00**	**45.11**

accurate performance evaluation, we repeated the experiments five times with our model and reported the average results.

Synpase for multi-organ CT segmentation. This dataset comprises 30 abdominal CT scan samples, and 18 were used for training and 12 for evaluation according to the method used by nnFormer [17].

ACDC for automated cardiac diagnosis. This dataset comprises cardiac MRI images of 100 patients. Following the approach used by nnFormer [17], data were divided into 70 training images, 10 validation images, and 20 evaluation images.

B. Evaluation metrics

We used the Dice score and 95% Hausdorff Distance (HD95) to evaluate the performance of our model. The Dice score measures the overlap between the volumetric segmentation predictions and voxels of the ground truths, and is defined as follows:

$$\begin{aligned}\text{Dice score}(G,P) &= \frac{2*\text{TP}}{(\text{TP+FP})+(\text{TP+FN})} \\ &= \frac{2*\sum_{v=1}^{V} G_v \cdot P_v}{\sum_{v=1}^{V} G_v^2 + \sum_{v=1}^{V} P_v^2},\end{aligned} \tag{8}$$

where TP, FP, and FN represent true positive, false positive, and false negative, respectively. G and P represent the ground truth and prediction values for voxel *v*.

HD95 uses the 95^{th} percentile of the distance between boundaries of the volumetric segmentation predictions and the voxels of the ground truths, and is defined as follows:

$$\text{HD}(G', P') = \max\{\max_{g' \in G'} \min_{p' \in P'} d(g', p'), \max_{p' \in P'} \min_{g' \in G'} d(g', p')\}, \tag{9}$$

where G' and P' represent the ground truth and prediction surface points set, and *d* represents a distance operation.

C. Implementation details

We conducted our experiments on Python 3.8, PyTorch 1.11, and CentOS7. We used an NVIDIA GeForce RTX 3090 for the experiment. During the experiments on the Synapse dataset, the epoch, iteration, and batch size were set to 1000, 250, and 2, respectively. For the experiments on the ACDC dataset, the epoch, iteration, and batch size were set to 1000, 250, and 4, respectively. We used stochastic gradient descent with a momentum of 0.99 as the optimizer. We used PolynomialLR as the learning rate scheduler, and the initial learning rate and weight decay were set to 0.01 and 3e-5, respectively. We followed the preprocessing method proposed by nnFormer [17] for the Synapse and ACDC datasets, and did not use any additional data. Both datasets used cropped images of size 128 $\times$ 128 $\times$ 64 and 160 $\times$ 160 $\times$ 16, respectively, with applied data augmentation techniques. The resolution of the cropped images is set to be consistent with the experiments conducted on nnFormer [17], ensuring experimental consistency.

Additionally, multiple resolution losses were computed in the training process according to the deep supervision scheme. The patch sizes used in patch partition were (4, 4, 2) and (4, 4, 1). The number of heads in the multi-head attention is 4. The window sizes used in the Swin transformer were (4, 4, 4) and (5, 5, 2), respectively.

IV. Results

A. Comparison with state-of-the-art models

Table I shows the Dice score and HD95 results of our proposed model compared to the state-of-the-art models on the abdominal multi-organ Synapse dataset. We improved the average Dice score by 5.86% (from 90.51% to 95.81%) for the spleen, 1.47% (from 86.25% to 87.52%) for the right kidney, 0.85% (from 86.57% to 87.31%) for the left kidney, and 6.38% (from 70.17% to 74.65%) for the gallbladder, compared to the latest state-of-the-art model (nnFormer). We also obtained the second-best results for the liver, stomach, aorta, and pancreas. Our proposed model outperformed nnFormer by 0.97% (from 86.57% to 87.41%) in terms of the average Dice score and 7.43% (from 10.63mm to 9.84mm) in terms of HD95. Furthermore, our proposed model demonstrated the efficiency of our module structure by reducing the space and time complexity of the model by 80.73% (from 150.50M to 29.00M) in parameters and 78.86% (from 213.40G to 45.11G) in FLOPS, compared to nnFormer.

We also trained our model on the ACDC dataset without data augmentation or pre-training. As a result, we achieved a 1.12% (from 90.94% to 91.96%) improvement compared to the state-of-the-art nnFormer model in right ventricle segmentation, as shown in Table II.

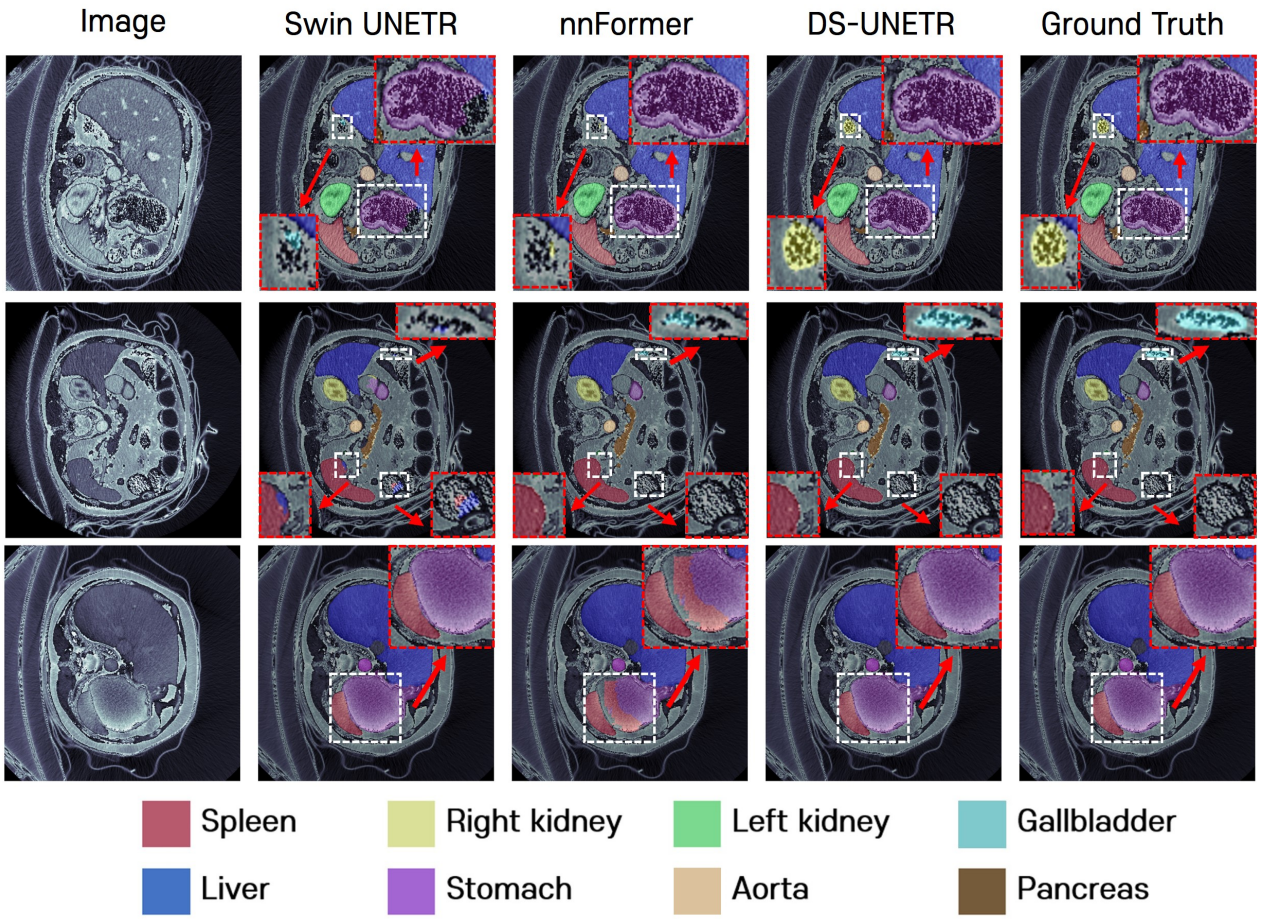


Fig. 4: Comparison of 2D multi-organ image segmentation results between Swin UNETR, nnFormer, and our proposed model (DS-UNETR)

TABLE II: Comparison on automatic cardiac diagnosis (ACDC). The abbreviations are: RV: Right Ventricle, Myo: Myocardium, LV: Left Ventricle. Best results are **bolded**. Best seconds are underlined.

Methods	RV	Myo	LV	Average Dice score	Params (M)	FLOPS (G)
TransUNet [12]	88.86	84.54	95.73	89.71	-	-
Swin-Unet [14]	88.55	85.62	**95.83**	90.00	-	-
UNETR [1]	85.29	86.52	94.02	88.61	92.69	33.83
MISSFormer [21]	89.55	88.04	94.99	90.86	-	-
nnFormer [17]	90.94	**89.58**	95.65	**92.06**	37.16	47.73
DS-UNETR	**91.96**	87.06	94.32	91.16	**28.06**	**25.67**

B. Qualitative analysis on the Synapse dataset

The multi-organ segmentation results of Swin UNETR, nnFormer, and our proposed model are visualized as 2D images in Fig. 4. The white dotted boxes in the images indicate the regions where each model segmented the organs, and the red dotted boxes, which are pointed to by arrows from the white dotted boxes, represent the enlarged images of the corresponding areas.

The first row of the image compares the segmentation results for the stomach (enlarged red dotted box at the top right of each image) and the right kidney (enlarged red dotted box at the bottom-left of each image). Swin UNETR failed to segment the right area of stomach properly and did not detect smaller organs such as the right kidney. The nnFormer segmented the stomach properly, but as in Swin UNETR, it failed to segment the right kidney. By contrast, our proposed model exhibited ground truth-like segmentation results for both the stomach and right kidney.

The second row shows the segmentation results of the gallbladder (enlarged red dotted box in the upper-left of each image) and spleen (enlarged red dotted box in the lower-left of each image). The area in the enlarged red box in the lower-right originally contained no organs, but it is shown to indicate the erroneous segmentation result of Swin UNETR.

TABLE III: Ablation study of DS-UNETR with Swin-Unet [14] and Swin UNETR [16].

Methods	Swin transformer in encoder	Swin transformer in decoder	DS-AE	Bi-ASF module	Average Dice Score on the Synapse dataset
Swin-Unet [14]	√	√			79.13
Swin UNETR [16]	√				83.48
DS-UNETR (w/o DS-AE & Bi-ASF)	√	√			84.10
DS-UNETR (w/o DS-AE)	√	√		√	85.10
DS-UNETR (w/o Bi-ASF)	√	√	√		86.68
DS-UNETR	√	√	√	√	**87.41**

Swin UNETR failed to capture the gallbladder entirely, and incorrectly segmented the upper right end of the liver as the kidney. The nnFormer failed to properly capture the gallbladder. Our proposed model showed the most similar results to the ground truth for the gallbladder and spleen.

The third row shows the segmentation results of the stomach and spleen (enlarged red dotted box in the upper-right of each image). The spleen is located in the lower left area of the red dotted box. The nnFormer exhibited a false positive, where the segmented area of the spleen partially overlapped with the stomach. However, Swin UNETR and our proposed model successfully segmented the stomach and spleen.

Additionally, the multi-organ segmentation results of Swin UNETR and our proposed model are visualized as 3D images in Fig. 5. The white dotted boxes in the images indicate the regions where each model segmented the organs. The red dotted boxes, which are pointed to by arrows from the white dotted boxes, represent the enlarged images of the corresponding areas.

The first row compares the segmentation results of the left kidney (enlarged red dotted box at the bottom-left of each image) and pancreas (enlarged red dotted box at the top-right of each image). Swin UNETR segmented the kidney well, but there was a small protruding area (bright green) on the left that was not present in the ground truth. Additionally, a small area in the pancreas was mislabeled in front of the stomach area. The second row compares the liver segmentation results (enlarged red dotted box at the top-left of each image). Swin UNETR showed incorrect segmentation results where the gallbladder invaded part of the liver area. By contrast, our proposed model showed good segmentation results similar to the ground truth for all three organs: the left kidney, pancreas, and liver.

C. Ablation study

Table III presents the results of the ablation study, which demonstrate the effectiveness of our proposed model.

Swin-Unet uses Swin transformer in both the encoder and decoder, but it is a 2D image-based framework that does not consider the depth (D). Therefore, it cannot utilize the relational information between each slice of a 3D image. Additionally, while Swin UNETR considers the depth for 3D image segmentation, the use of the CNN in the decoder prevents it from sufficiently capturing global information of the feature map.

We used the attentions that consider the depth of the 3D image in both the encoder and decoder as the basic structure

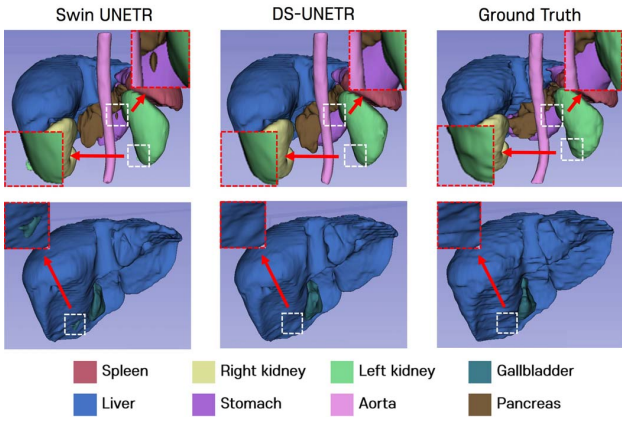


Fig. 5: Comparison of 3D multi-organ image segmentation results between Swin UNETR and our proposed model (DS-UNETR).

(without the DS-AE and Bi-ASF module). As a result, we achieved a 6.28% improvement in terms of the average Dice score compared to Swin-Unet (from 79.13% to 84.10%), and 0.74% improvement compared to Swin UNETR (from 83.48% to 84.10%).

To demonstrate the importance of the DS-AE and Bi-ASF module, we examined the changes in average Dice score when each function was added to our basic structure. First, when only the Bi-ASF module that fuses the features of all scales was added, we achieved a 1.19% improvement (from 84.10% to 85.10%) compared to our basic structure. When only the DS-AE, which adds inter-channel relation information, was applied, we achieved a 3.07% improvement (from 84.10% to 86.68%) compared to our basic structure. As a result, we found that effectively learning the importance of channels using the DS-AE further improved 3D image segmentation performance of our proposed model. Finally, our proposed model, which applies both the DS-AE and Bi-ASF module, achieved the best performance by improving the average Dice score by 3.93% (from 84.10% to 87.41%). It is worth noting that trained the model without pretraining or additional data.

V. Conclusion

We propose DS-UNETR for 3D medical image segmentation, which comprises the DS-AE with spatial attention stream, channel attention stream, and the fusion module, and the Bi-ASF module that combines attention information extracted

from all stages. The DS-AE effectively captures both spatial and channel features in parallel streams to better understand the spatial and channel relationships in 3D medical images. This makes it useful for learning contextual information across various structures. Furthermore, the Bi-ASF module enhances 3D medical image segmentation performance by effectively transmitting all local and global information generated in the encoder to the decoder through the fusion of attention information at all scales. According to our experiments, our proposed model, DS-UNETR, demonstrates superior performance and efficiency not only in terms of segmentation accuracy, but also considering the model complexity (both spatial and temporal aspects) compared to existing state-of-the-art models on 3D medical image segmentation benchmark datasets. We believe that our proposed model can be effectively utilized in various applications for medical big data analysis.

ACKNOWLEDGEMENT

This work was supported by the Institute of Information and Communications Technology Planning and Evaluation (IITP) Grant by the Korean Government through Ministry of Science and ICT (MSIT) (IITP-2017-0-00477, (SW Star Lab) Research and Development of the high performance in-memory distributed DBMS based on flash memory storage in an IoT environment) and (No. 2020-0-01361, Artificial Intelligence Graduate School Program (Yonsei University)).

REFERENCES

[1] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu, "Unetr: Transformers for 3d medical image segmentation," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2022, pp. 574–584.

[2] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature methods*, vol. 18, no. 2, pp. 203–211, 2021.

[3] F. Ramzan, M. U. G. Khan, S. Iqbal, T. Saba, and A. Rehman, "Volumetric segmentation of brain regions from mri scans using 3d convolutional neural networks," *IEEE Access*, vol. 8, pp. 103 697–103 709, 2020.

[4] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.

[5] H. Hu, Z. Zhang, Z. Xie, and S. Lin, "Local relation networks for image recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3464–3473.

[6] C. Chen, X. Liu, M. Ding, J. Zheng, and J. Li, "3d dilated multi-fiber network for real-time brain tumor segmentation in mri," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part III 22*. Springer, 2019, pp. 184–192.

[7] A. Lin, B. Chen, J. Xu, Z. Zhang, G. Lu, and D. Zhang, "Ds-transunet: Dual swin transformer u-net for medical image segmentation," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–15, 2022.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[9] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of naacL-HLT*, vol. 1, 2019, p. 2.

[10] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2020.

[11] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[12] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[13] Y. Zhang, H. Liu, and Q. Hu, "Transfuse: Fusing transformers and cnns for medical image segmentation," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24*. Springer, 2021, pp. 14–24.

[14] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-unet: Unet-like pure transformer for medical image segmentation," in *European conference on computer vision*. Springer, 2022, pp. 205–218.

[15] X. Chen, X. Wang, K. Zhang, K.-M. Fung, T. C. Thai, K. Moore, R. S. Mannel, H. Liu, B. Zheng, and Y. Qiu, "Recent advances and clinical applications of deep learning in medical image analysis," *Medical Image Analysis*, vol. 79, p. 102444, 2022.

[16] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, and D. Xu, "Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images," in *International MICCAI Brainlesion Workshop*. Springer, 2021, pp. 272–284.

[17] H.-Y. Zhou, J. Guo, Y. Zhang, X. Han, L. Yu, L. Wang, and Y. Yu, "nnformer: volumetric medical image segmentation via a 3d transformer," *IEEE Transactions on Image Processing*, 2023.

[18] U. Baid, S. Ghodasara, S. Mohan, M. Bilello, E. Calabrese, E. Colak, K. Farahani, J. Kalpathy-Cramer, F. C. Kitamura, S. Pati *et al.*, "The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification," *arXiv preprint arXiv:2107.02314*, 2021.

[19] M. Tan, R. Pang, and Q. V. Le, "Efficientdet: Scalable and efficient object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 781–10 790.

[20] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[21] X. Huang, Z. Deng, D. Li, X. Yuan, and Y. Fu, "Missformer: An effective transformer for 2d medical image segmentation," *IEEE Transactions on Medical Imaging*, 2022.

[22] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 fourth international conference on 3D vision (3DV)*. Ieee, 2016, pp. 565–571.