IEEE BigComp2024

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

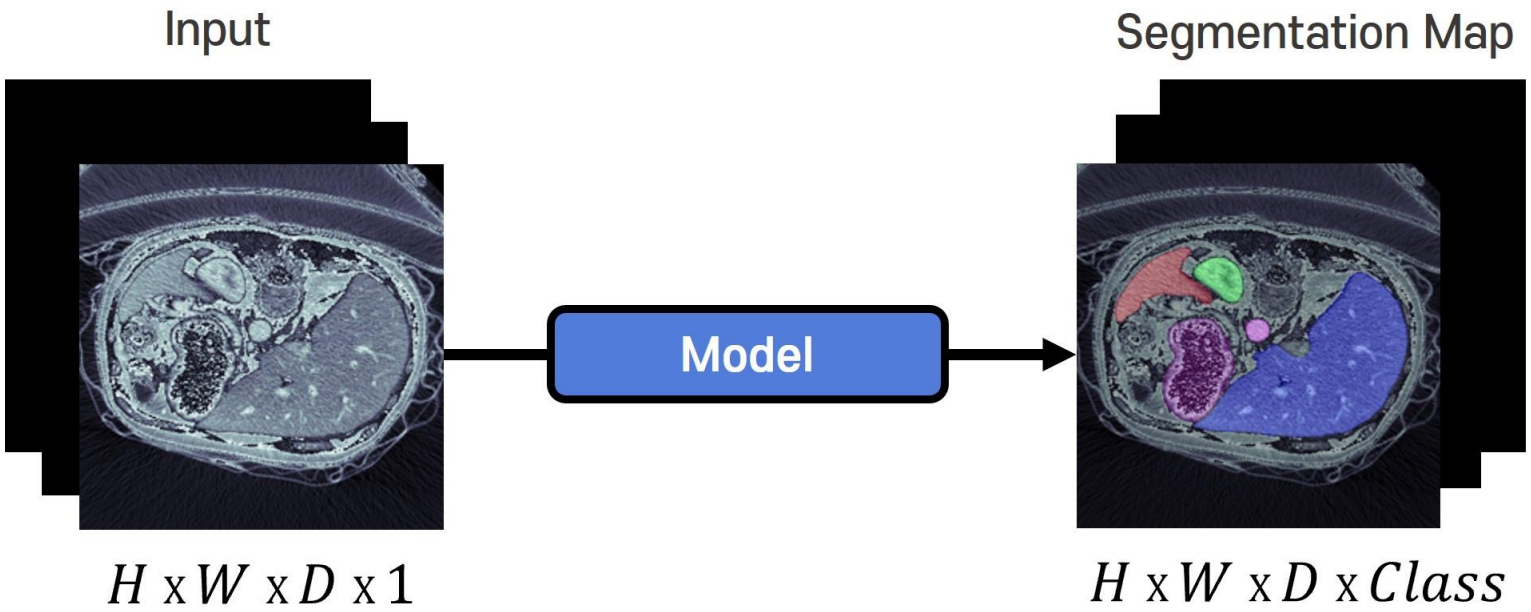
{highsk, skd, jyy1551, sanghyun}@yonsei.ac.kr

* Equal contribution
† Corresponding author

Introduction

3D Medical Image Segmentation

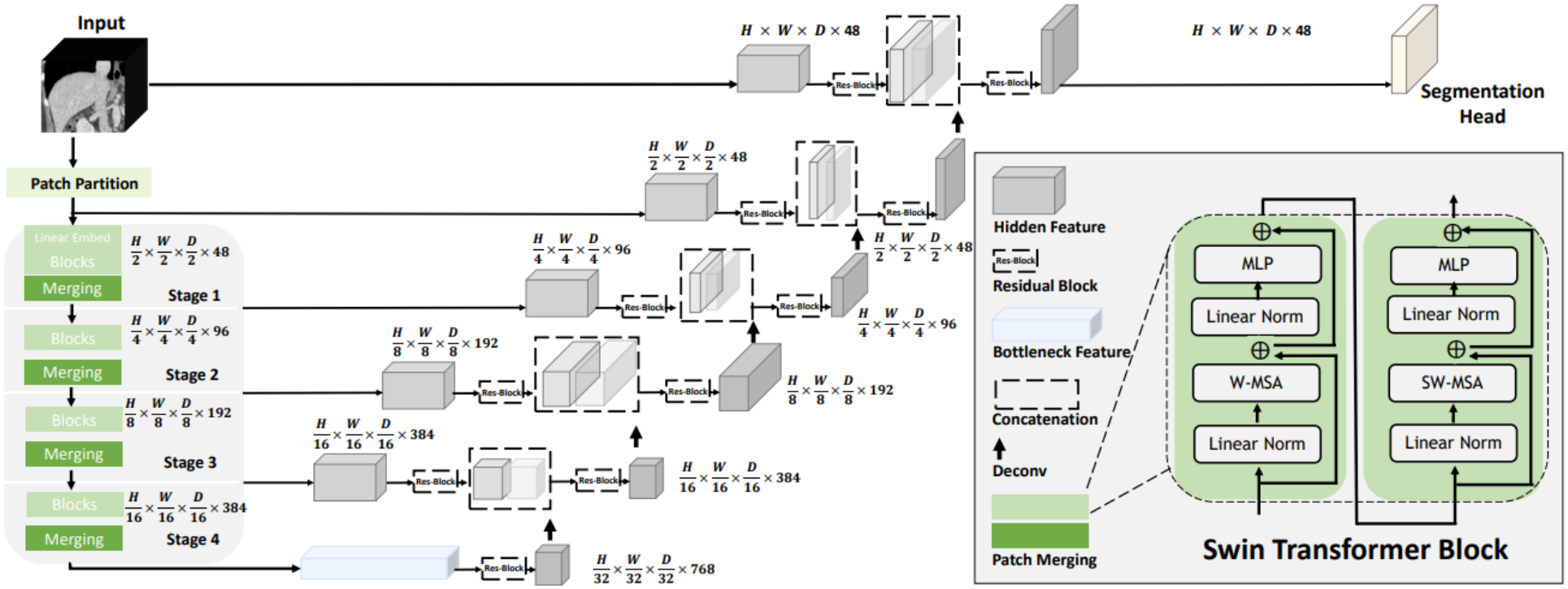
- Distinguishing lesions, tumors, and organs in complex 3D CT or MRI images
- Supervised learning is utilized to create a **segmentation map** indicating tissue location for each pixel
- Utilized in medical diagnostics, surgical support, disease research, demanding accuracy and promptness



Introduction

Swin UNETR

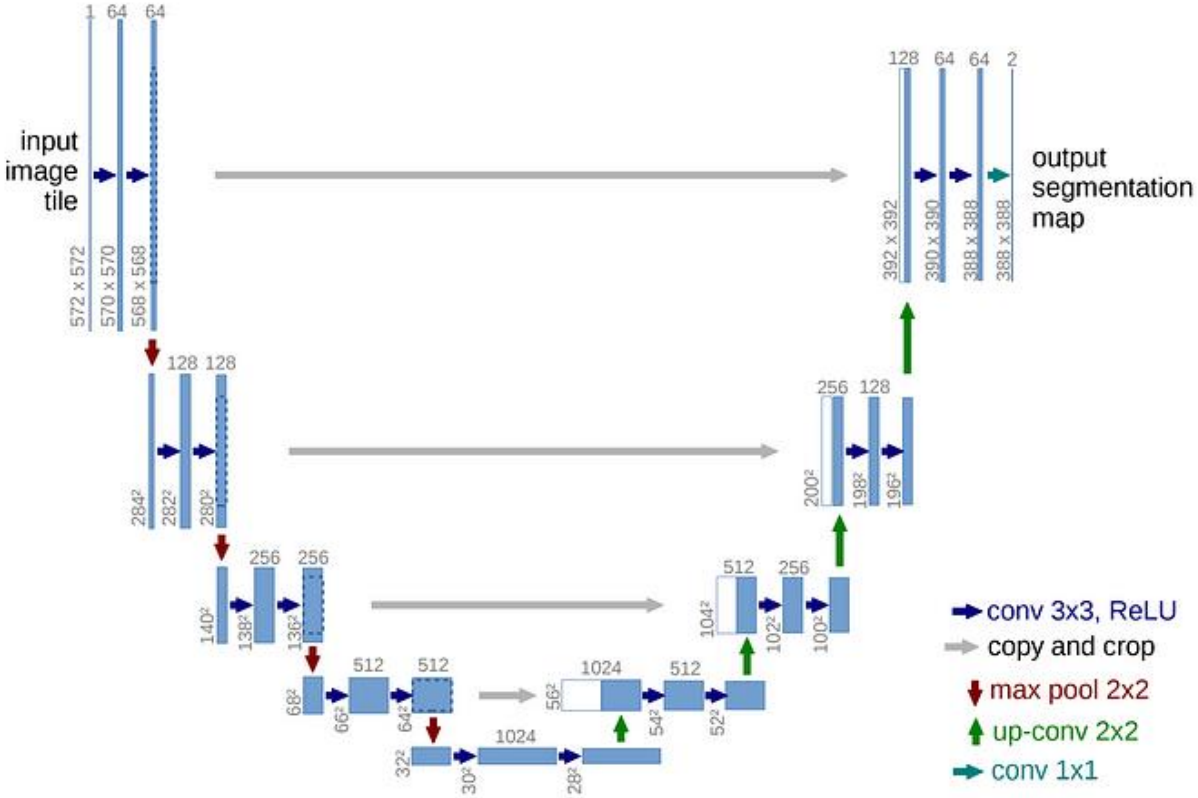
- U-shaped network, comprising an encoder and a decoder, is commonly employed in segmentation models
- **Encoder** captures multi-scale information, **Decoder** integrates it to produce the segmentation map
- Swin UNETR, utilizing Swin Transformer in the encoder, has gained prominence recently

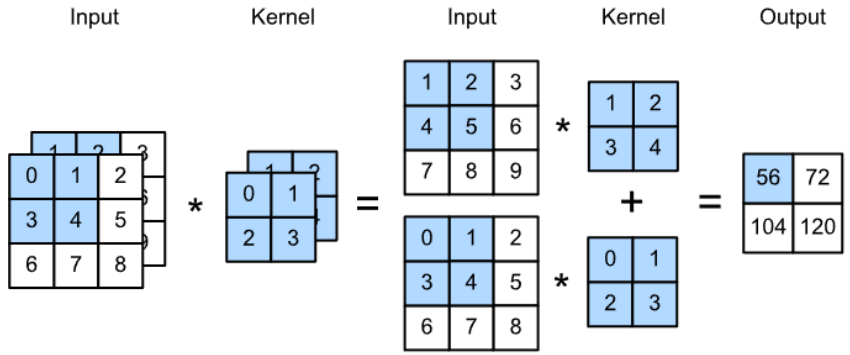


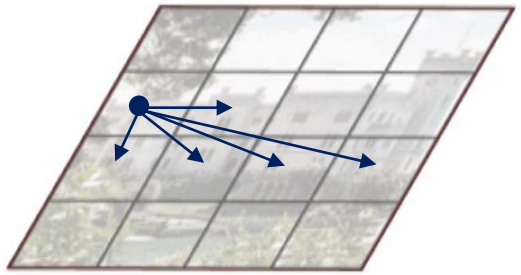
Tang, Yucheng, et al. "Self-supervised pre-training of swin transformers for 3d medical image analysis." CVPR 2022

Limitations of Traditional Approaches

- Increasing channels in the encoder produces diverse contextual information, but focuses only on spatial features
- Features from both the upper and lower stages are combined only once to acquire the local and global information necessary for segmentation



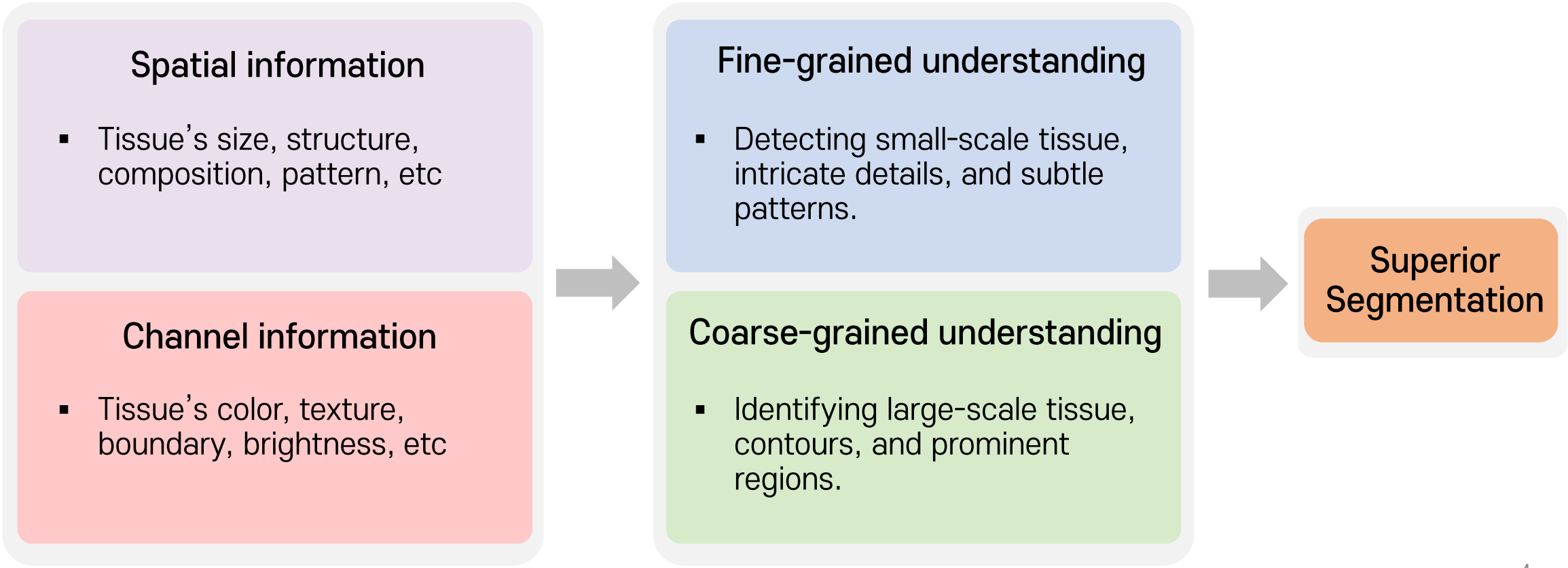




Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." MICCAI 2015

Contributions

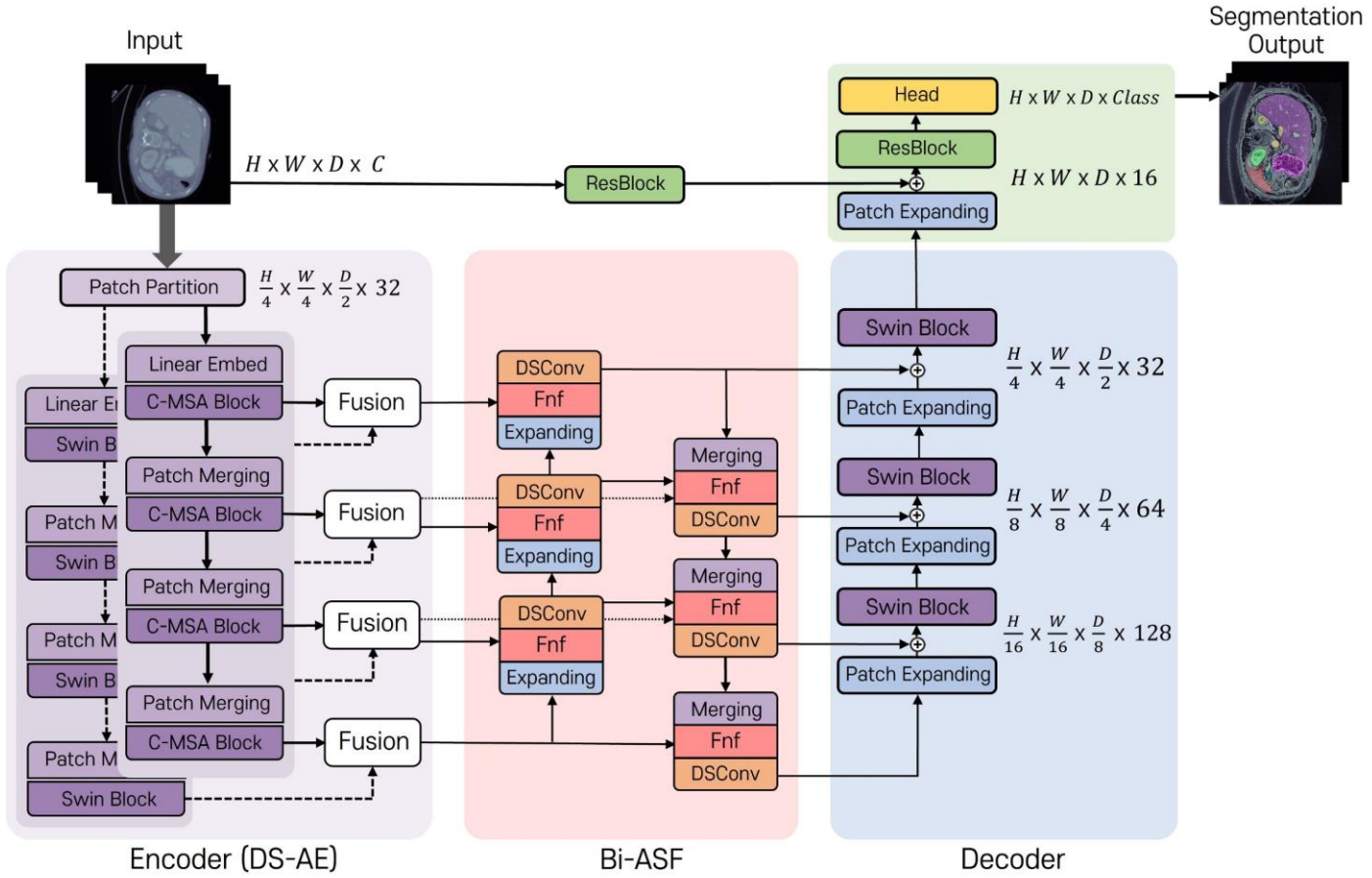
- Through dual attention, useful spatial and channel information is acquired at each stage of the model
- Fusing various scales of information repeatedly, enabling fine and coarse-grained understanding
- Structuring the model efficiently, decreasing the time and space complexity!



Method

Dual Stream Fusion U-Net Transformers

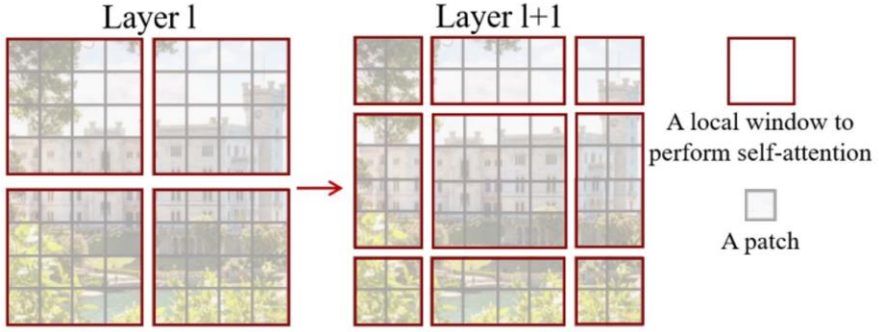
- **DS-AE** performs spatial attention and channel attention in parallel and then fuses them together
- **Bi-ASF** effectively integrates local and global information
- **DS-UNETR** is a model constructed with a U-shaped network using two designed modules



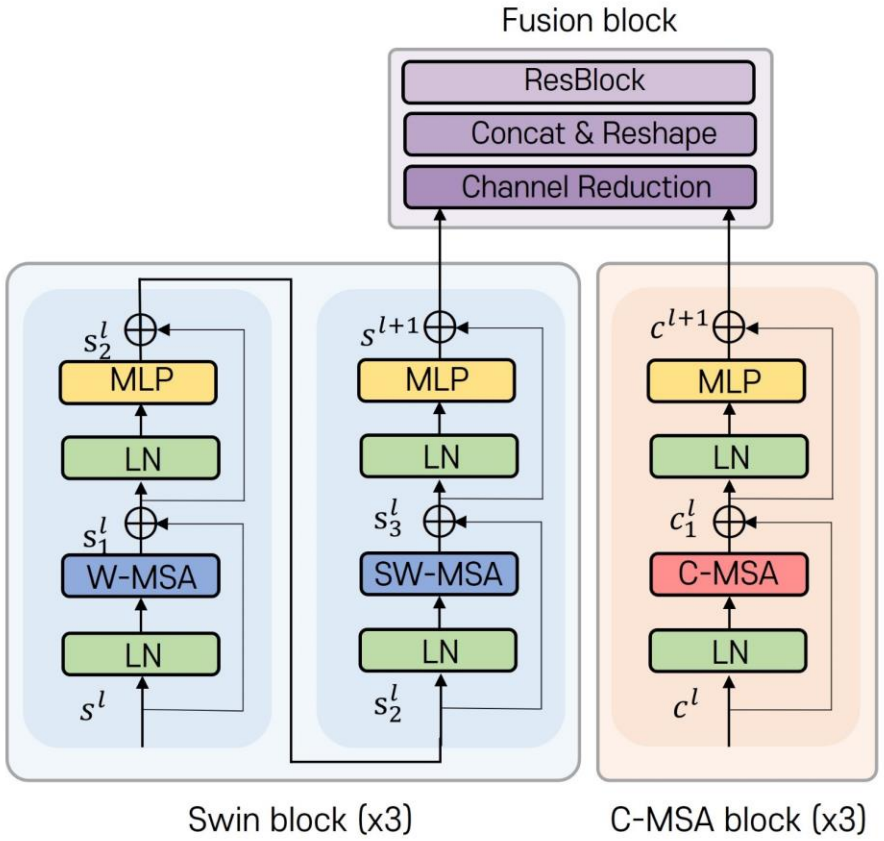
Method

Dual Stream Attention

- **Swin block** conducts spatial attention to learn spatial relationships, acquiring spatial information
- W-MSA: apply local attention for *O* (token) complexity!
- SW-MSA: shift the window after each layer to transfer information between local windows



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

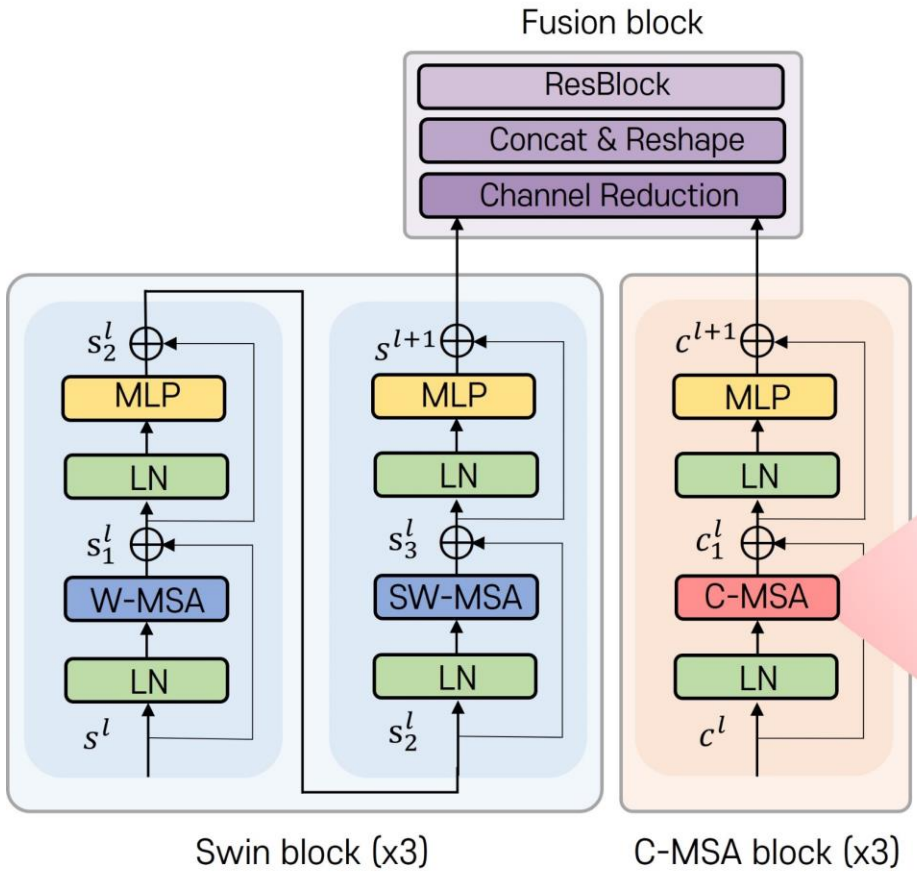


Liu, Ze, et al. "Swin transformer: Hierarchical vision transformer using shifted windows." ICCV 2021

Method

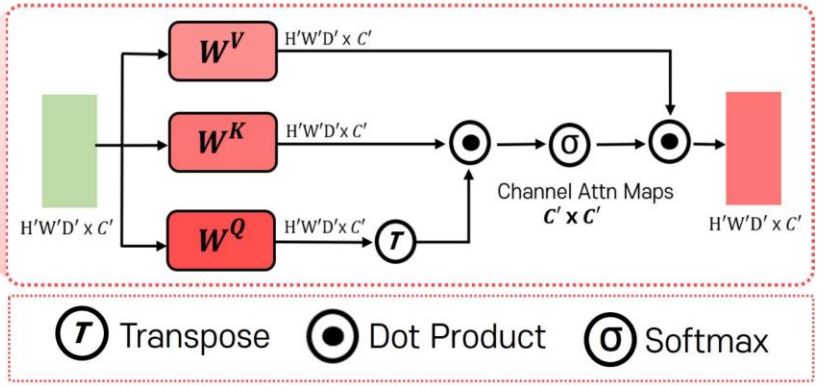
Dual Stream Attention

- **C-MSA block** conducts channel attention to learn channel relationships, acquiring contextual information
- Channel Attention: generate channel attention map and consider only relevant channels
- C-MSA: perform Channel Attention divided into multiple channels, improving efficiency!



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

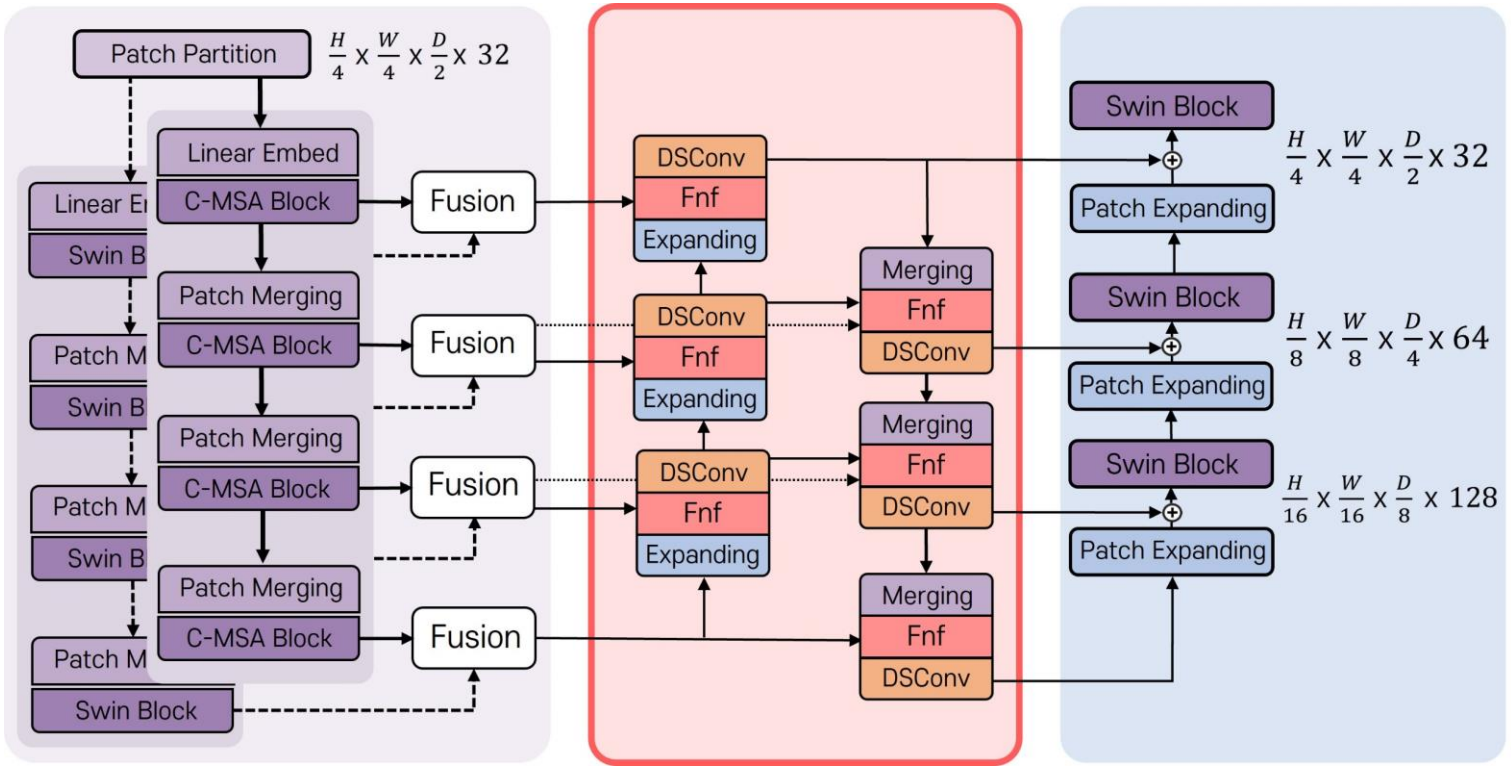
$$\text{Channel Attention}(Q, K, V) = V \cdot \text{Softmax}(\frac{Q^T K}{\sqrt{d}}),$$



Method

Bidirectional All Scale Fusion

- Bidirectionally fusing features of various scales, effectively leveraging both local and global information
- Fnf: simple fusion method for combining many feature maps with different resolutions
- Fused features are recalibrated through DSConv



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

$$F_3^{bu} = \text{DSConv}\left(\frac{w_1 \cdot z^9 + w_2 \cdot \text{Resize}(z^{12})}{w_1 + w_2 + \epsilon}\right),$$

$$F_3^{td} = \text{DSConv}\left(\frac{w_1' \cdot z^9 + w_2' \cdot F_3^{bu} + w_3' \cdot \text{Resize}(F_2^{td})}{w_1' + w_2' + w_3' + \epsilon}\right),$$

Tan, Mingxing, Ruoming Pang, and Quoc V. Le. "Efficientdet: Scalable and efficient object detection." CVPR 2020

Results

Multi organ Image segmentation

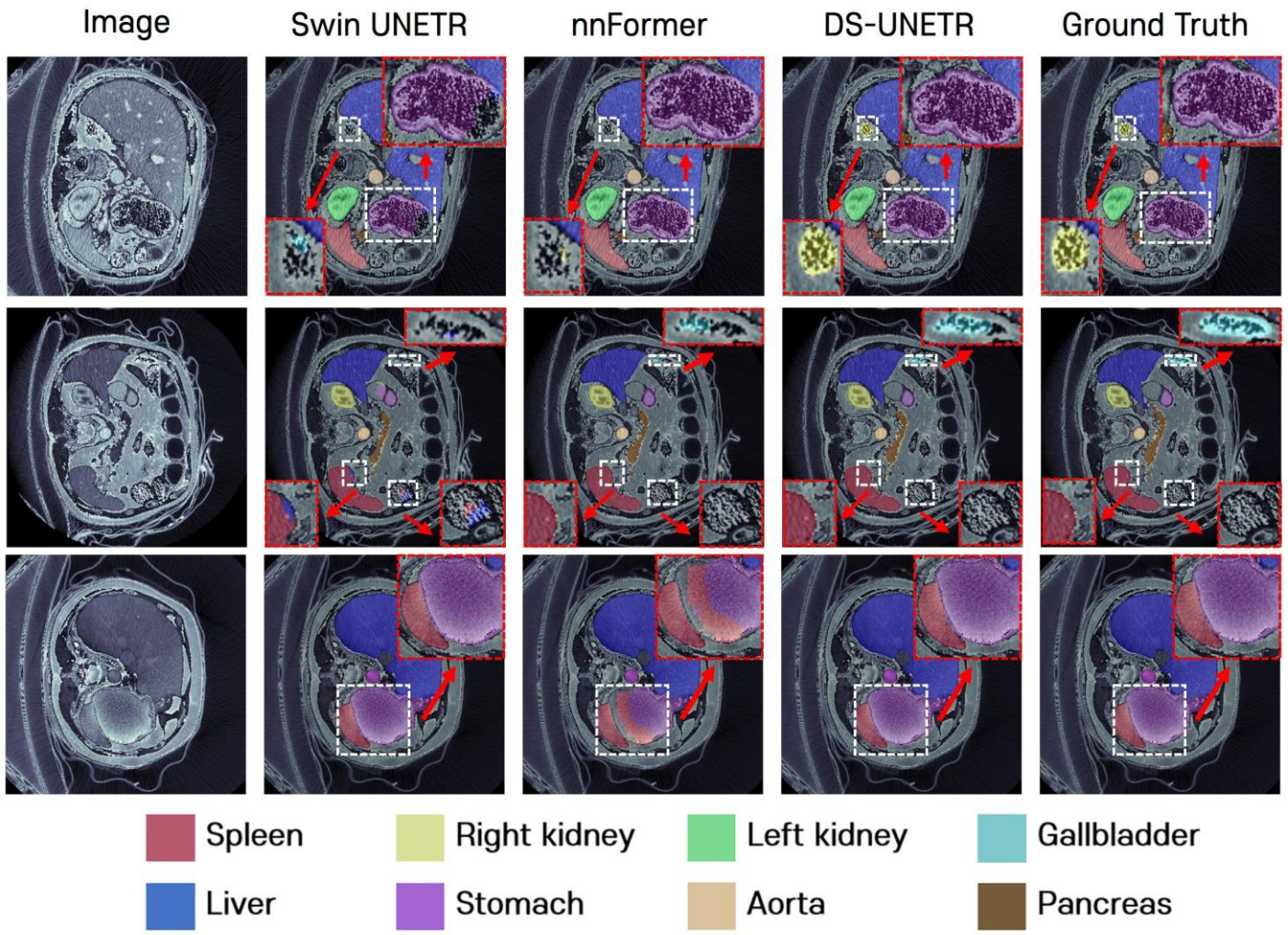
- Dice Score: A metric that combines precision and recall into a single measure
- HD95: A method to determine distance between two sets
- Proposed model achieved SOTA performance in average Dice Score, with the smallest Params and FLOPS

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**

Results

Multi organ Image segmentation



- High segmentation performance for the stomach, accurately identifies the right kidney
- Does not overlap with other organs, and accurately identifies the gallbladder
- Distinguish between Spleen and Stomach well without confusion

Results

Cardiac Image Segmentation

- Achieved SOTA performance for Right Ventricle!
- Average Dice Score achieves the second best performance
- Commendable performance with fewer parameters and computational FLOPS compared to nnFormer

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**

Results

Ablation study

- 3D Swin Transformer in both the encoder and decoder of U-Net is utilized in the basic structure
- Using the Bi-ASF module improves performance by 1% by better leveraging local and global information
- Utilizing the DS-AE module improves performance by 2.58% through better utilization of contextual information

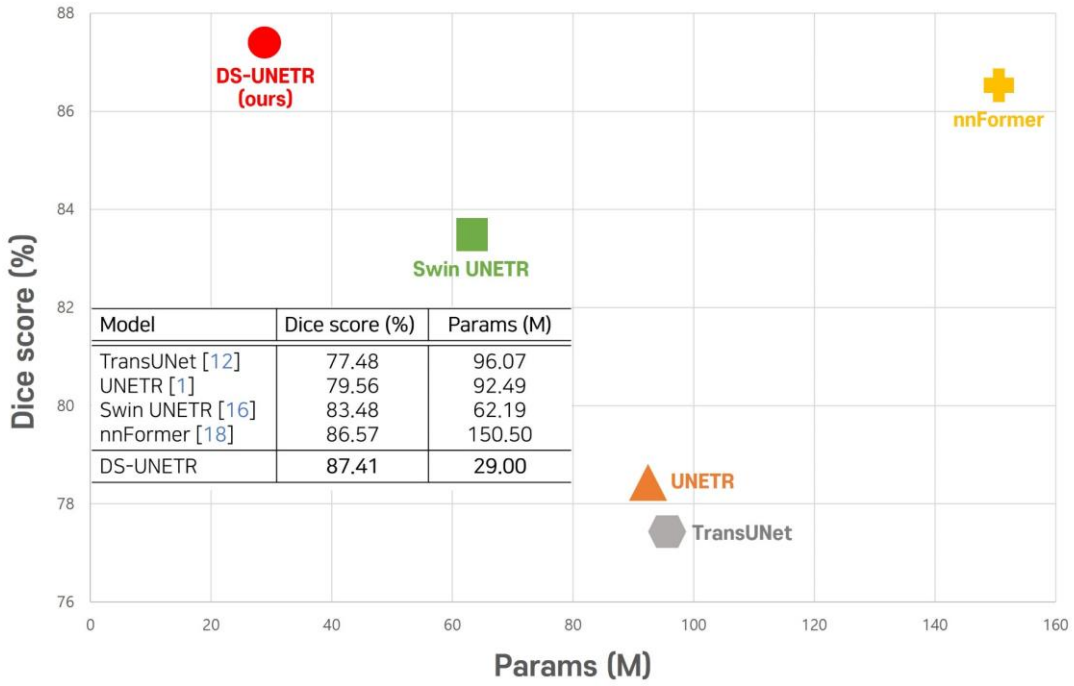
TABLE III: Ablation study of DS-UNETR with Swin-Unet [14] and Swin UNETR [16].

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**

Conclusions

DS-UNETR

- To enhance segmentation, propose **effective utilization of channel information** and **feature fusion** methods
- Proposed model enables accurate and efficient medical image segmentation
- Expect the proposed model to be utilized in various medical big data analyses!



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

Thank you