ECAI 2024
U. Endriss et al. (Eds.)

doi:10.3233/FAIA240505

Dual Attention Encoder with Joint Preservation for Medical Image Segmentation

Shijie Li[a], Yunbin Tu[b], Yu Gong[a], Bowen Zhong[a] and Zheng Li[a,*]

[a]College of Computer Science, Sichuan University
[b]School of Computer Science and Technology, University of Chinese Academy of Sciences

Abstract. Transformers have recently gained considerable popularity for capturing long-range dependencies in the medical image segmentation. However, most transformer-based segmentation methods primarily focus on modeling global dependencies and fail to fully explore the complementary nature of different dimensional dependencies within features. These methods simply treat the aggregation of multi-dimensional dependencies as auxiliary modules for incorporating context into the Transformer architecture, thereby limiting the model's capability to learn rich feature representations. To address this issue, we introduce the Dual Attention Encoder with Joint Preservation (DANIE) for medical image segmentation, which synergistically aggregates spatial-channel dependencies across both local and global areas through attention learning. Additionally, we design a lightweight aggregation mechanism, termed Joint Preservation, which learns a composite feature representation, allowing different dependencies to complement each other. Without bells and whistles, our DANIE significantly improves the performance of previous state-of-the-art methods on five popular medical image segmentation benchmarks, including Synapse, ACDC, ISIC 2017, ISIC 2018 and GlaS.

1 Introduction

Medical image segmentation is one of the key tasks in computer vision, which provides valuable information about the anatomy areas that are needed for the detailed analysis. Accurate segmentation enables physicians to obtain dependable morphological statistics, which is critical for disease diagnosis. Different from conventional images, medical images usually contain intricate tissue structures and blurred edges,which pose a huge challenge to efficiently segment specific targets from medical images.

Inspired by Fully Convolutional Network (FCN) [19] that boosts the performance of semantic segmentation, Convolutional Neural Networks (CNNs) have been the de-facto standard for medical image analysis tasks. For instance, Unet [22] tries to apply fully convolutional networks to medical image segmentation, where it employs an encoder to extract features and a symmetric decoder to gradually upsample spatial dimensions. This classic paradigm has been followed by some methods [23, 39], due to its simple structure and efficient performance. However, the fixed receptive field of convolution operators hinders these CNN-based methods to capture long-range relationships between distant pixels in medical images. Recently, motivated by the success of vision transformers in natural image segmentation, the transformer-based segmentation methods [7, 5] have been developed to address the drawbacks of CNNs in medical image segmentation. These methods leverage the global relation modeling abilities of transformers to interact long-distance pixel information, thereby identifying key features across the entire input.

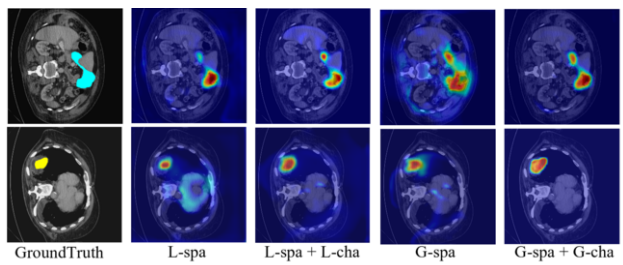

Figure 1. Visualization of class activation maps in different structures with Grand-CAM [24]. L-spa/cha is local spatial/channel attention; G-spa/cha denotes global spatial/channel attention.

Despite transformer-based segmentation methods have shown the progress, they fail to make full use of the dependencies between spatial and channel dimensions within features. This results in an inadequate understanding of feature semantics during target segmentation. Specifically, these methods [38, 21] focus on learning global dependencies through self-attention, which may overlooks the complementary nature of dependencies across different dimensions. That is, spatial dependencies involve the relationships between adjacent pixel regions, which help to recognize the location of targets (such as tissue boundaries, texture, etc.). Besides, learning channel-wise dependencies can highlight the important feature semantics across various channels, thus revealing tissue structures that are not apparent in a single channel. In short, both spatial and channel dependencies not only play its unique role in improving feature representation, but also supplement each other. Therefore, synergistically learning both dependencies can enhance the feature diversity in the spatial dimension and improve feature utilization in the channel dimension, so as to boost the model's capability for localizing anatomical structures. As shown in Figure 1, class activation maps from the structures aggregating both dimensional dependencies can better target objects than those aggregating dependencies along a single dimension, either on a global or local scale.

In this paper, we propose a **D**ual **A**ttention E**N**coder with Jo**I**nt Pr**E**servation (DANIE) architecture that synergistically aggregates spatial-channel dependencies in both local and global areas, to construct the rich feature representation of medical image. Architecture-

* Corresponding Author. Email: lizheng@scu.edu.cn

wise, given two feature maps extracted from a medical image, we design a Dual Attention Encoder (DAE) to capture their spatial-channel dependencies at multi-scales (*e.g.,* four stages). At the first stage, the encoder uses a *self-attention perception* (SAP) to perceive the global spatial dependencies of one feature map via interacting its all regions, alongside a *hierarchical-attention perception* (HAP) to dynamically model the channel and spatial correlations of the other feature map within local areas. In the following stages, we repeat the above procedure to progressively model the multi-scale feature maps with spatial-channel dependencies.

Next, the obtained feature maps are fed into the Joint Preservation (JP) to learn a composite representation that encapsulates global and local dependencies across spatial-channel dimensions. Specifically, JP contains two light-weight components: the *enhancing self-attentional perceptron* is developed to learn channel representations for global spatial dependencies from the feature maps of SAP, and the *enhancing hierarchical-attention perception* is devised to enhance non-local interactions for local spatial and channel feature from the feature maps of HAP. Finally, the fused feature map is used to generate the segmentation image.

Our contributions can be summarized as follows:

- We propose a novel DANIE that aggregates global and local spatial-channel features, in order to obtain a powerful composite representation. Meanwhile, by leveraging these fine-grained local dependency features at a larger scale, the model achieves a deeper comprehension of intricate details.
- In DANIE, we first design DAE to not only capture global spatial information, but also learn spatial-channel dependencies within local regions. Then, JP is devised to learn a composite features representation, where it integrates multiple dimensional dependencies within a unified block and ensure these dependencies mutually supplement each other.
- We demonstrate the effectiveness of DANIE on five challenging benchmark datasets: Synapse, ACDC, Skin Lesion segmentation (ISIC 2017, ISIC 2018) and GlaS. Our DANIE produces better results than state-of-the-art methods. Notably, our DANIE outperforms recent transformer-based methods CASCADE [21] (85.47% vs. 82.68%) using fewer computational cost (26.55 GFLOPs vs. 44.83 GFLOPs) on Synapse multi-organ dataset.

2 Related Work

2.1 CNN-based Segmentation Method

UNet [22] pioneers the application of CNNs for medical image segmentation and shows promising progress. Inspired by this, numerous subsequent methods have adopted the U-shaped, fully convolutional network (FCN) design [39]. For instance, UNet++ [39] introduces dense skip connections that link the encoder and decoder subnetworks. The nnUNet [16], a highly influential variant of the UNet architecture, uniquely automates data preprocessing and the selection of optimal network architectures tailored for specific tasks. Recently, Han et al. [13] develop a 2.5D 24-layer FCN for liver segmentation tasks, integrating a residual block into the model to improve its performance. However, these CNN-based segmentation methods are limited in performance due to their fixed receptive field. To address this, we introduce global channel and spatial attention mechanisms to enhance the interaction of distant feature information.

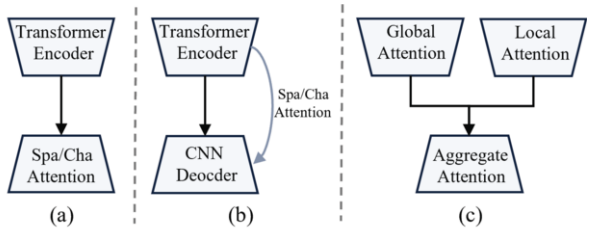

Figure 2. Conceptual comparison of recent Transformer-based approaches. (a) employs a Transformer as backbone, supplemented with attention mechanism as decoder to enhance multi-dimensional information [21]. (b) combines Transformer and CNN for feature extraction, using attention mechanisms to augment feature representation [10, 38]. (c) is our proposed efficient dual attention aggregate structure.

2.2 Transformer-based Segmentation Method

With the success of Transformers in computer vision [27, 29], the transformer-based methods for medical image segmentation become popular. TransUNet [7] tries to combine CNNs and Transformers, where CNNs play the role of local modeling and basic feature extraction, while Transformers can capture long-range dependencies. SwinUNet [5] designs a U-shaped architecture based on pure Swin Transformer blocks. SSFormer [33] introduces a layered architecture that leverages the original Swin Transformer as its backbone, achieving a reduction in parameters by merging image patches and employing shifted windows. Recent research investigates the incorporation of attention mechanisms as assisted modules with Transformers. For instance, TransFuse [38] proposes the BiFusion module based on attention mechanism for fusing Transformer and CNN branches to maximize the advantages offered by both. CASCADE [21] introduces an attention-based decoder combined with a Transformer encoder. Although existing transformer-based methods add attention mechanisms as auxiliary to enhance the expressive capability of the Transformer architecture, these models still fail to fully exploit the intrinsic correlations among features. In contrast, we analyze and synergistically utilize the complementarity of dependencies across different dimensions, thereby improving the perception of detail information.

2.3 Attention Mechanism

Attention Mechanism is widely adopted as an important ingredient in various vision tasks [28, 30]. For example, Hu et al. [15] design a squeeze-and-excitation block to compute channel attention and enhance important channel feature maps. Wang et al. [34] are inspired by traditional methods [4] and generalize the classical non-local operation into deep neural networks. Subsequently, GCNet [6] proposes query-independent attention maps, which further reduce the amount of computation and achieve excellent performance in quality. Recently, to leverage attention mechanism for medical image segmentation, AttnUnet [20] combines spatial attention with UNet for abdominal pancreas segmentation. PraNet [11] leverages a guide map generated by aggregating high-level features during the decoding phase, which enhances boundary refinement through collaboration with reverse attention mechanisms.

3 Method

In this section, we first conduct an analysis of previous work. Subsequent sections will detail the formulations of DANIE. Specifically,

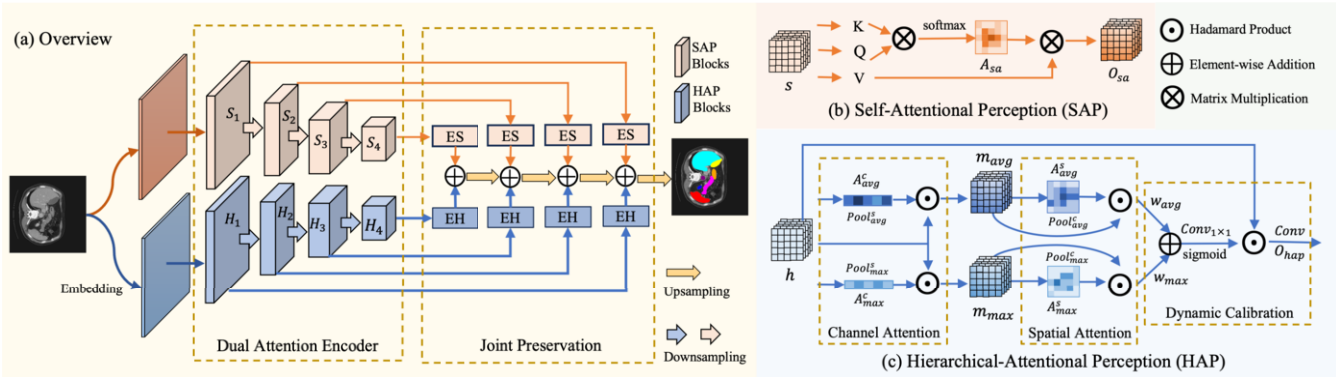

Figure 3. Our DANIE primarily consists of three parts: Embedding, Dual Attention Encoder (DAE) and Joint Preservation. (a) is overview of DANIE, (b) and (c) are the two attentional streams of the DAE.

our DANIE framework contains three key components: (a) *Embedding Layer*, (b) *Dual Attention Encoder* that model both global and local dependencies across spatial and channel dimensions, and (c) *Joint Preservation* for learning composite feature representation.

3.1 Analysis of Previous Work

As illustrated in Figure 2, while transformer-based methods can mitigate the limitations of CNN-based approaches and capture multi-scale information via attention modules, they still do not fully exploit the complementary interactions among spatial-channel dependencies. Essentially, these methods merely use attention mechanisms to augment the modeling capabilities of Transformer architectures. There are two key differences between our DANIE and previous works: (1) Fusion of global and local information. While prior studies such as [7, 38] have employed the fusion of global and local features, they typically rely on a hybrid structure of Transformers and CNNs for feature extraction, without selectively modeling feature representation. In contrast, our DANIE strategically activates global and local features from the perspective of spatial-channel dependencies interactions. (2) Aggregation of diverse dependencies. Unlike the coarse aggregation methods (e.g., element-wise addition, concatenation) employed in previous works, we analyse the property of various dependencies, and design a aggregate attention mechanism that effectively compensates for the deficiencies in extracting features.

3.2 Embedding Layer

The input images first pass through separate embedding layers to extract informative representations before fed into the Dual Attention Encoder. For the self-attentional perception (SAP) stream, the embedding layer divides the input image into multiple patches and learns a vector representation for each patch. When an image with dimensions $(H \times W \times 3)$ is input to the network, the embedding layer transforms the image into embedded patches $s \in \mathbb{R}^{(\frac{H}{4} \times \frac{W}{4} \times C)}$, where H, W, and C denotes the height, width, and number of channels respectively. The embedded patches is the input to SAP stream.

For the hierarchical-attentional perception (HAP) stream, the embedding layer applies convolutional filters across the input image to extract hierarchical features. When an image with dimensions $(H \times W \times 3)$ is input to the network, the embedding layer transforms it into feature maps $h \in \mathbb{R}^{(\frac{H}{4} \times \frac{W}{4} \times C)}$. The filters activate on low-level cues like edges and textures to generate C-channel feature representations, preserving local relationships in the image.

3.3 Dual Attention Encoder

We input the feature maps into a dual-stream module to learn multi-dimensional features. An independent self-attentional stream can learn the complete region-to-region correlation over a global range. The hierarchical attentional stream captures channel and spatial dimension dependencies, focusing on key local areas of the input.

Self-Attentional Perception. To capture global spatial relationships, we employ multi-head self-attention (MHA). MHA is an extension of the self-attention (SA) mechanism, where multiple SA blocks are applied in parallel. The outputs of these parallel SA blocks are concatenated and then projected back to the original dimension, resulting in a latent representation that encodes global spatial dependencies. Specifically, SA generates query, key and value vector representations for each input. The dot product between the query vectors and all key vectors yields an attention distribution, indicating the relevance of the other inputs as described in Figure 3 (b). A weighted sum of the value vectors is then computed based on the attention distribution to obtain a new contextualized representation. The formula for SA is:

$$\mathrm{SA}(Q, K, V) = softmax(QK^T)V \tag{1}$$

where Q, K, and V are the query, key and value matrices respectively. Parallelly computing multiple SA blocks and concatenating their outputs, MHA captures multi-scale interdependencies. Essentially, this self-attentional stream generates attention masks based on semantic information, highlighting crucial long-range spatial dependencies for segmenting complex organizational structures.

Hierarchical-Attentional Perception. Whilst self-attentional stream captures global dependencies of feature, it does not specify how to learn local detail in the channel and spatial dimensional. Therefore, we introduce a hierarchical-attentional stream alongside the self-attentional stream. This stream includes *channel attention* and *spatial attention* to highlight important features. Additionally, *dynamic calibration* is designed for the effective integration of these features, ensuring accurate localization of fine details.

Formally, given a feature map h, we compress its spatial dimensions via pooling and then apply a multi-layer perceptron (MLP) to derive channel attention maps: $A^c_{[max,avg]} \in \mathbb{R}^{C \times 1 \times 1}$, where max and avg denote the max pooling and average pooling branches, respectively. These attention maps are then applied to enhance the orig-

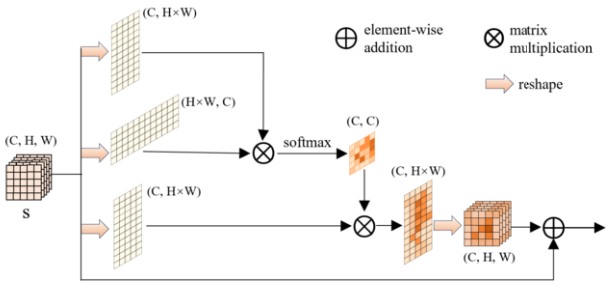

Figure 4. Details of the Enhancing Self-Attentional Perceptron (ES). The feature maps s are extracted from the SAP stream.

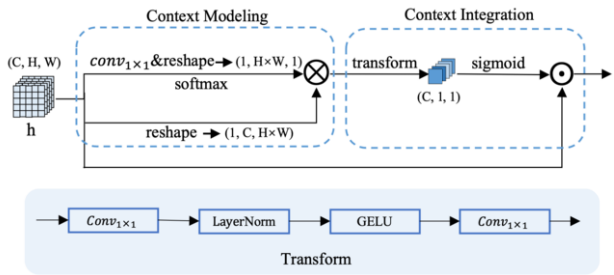

Figure 5. Details of the Enhancing Hierarchical-Attentional Perceptron (EH). The feature maps h are extracted from the HAP stream.

inal features, yielding weighted features: $m_{[max,avg]} \in \mathbb{R}^{H\times W\times C}$. Following this, we compress the channel dimension of the feature map $m_{[max,avg]}$ and then apply a convolutional layer to generate spatial attention maps: $A^s_{[max,avg]} \in \mathbb{R}^{1\times H\times W}$. These attention maps are then used to weight $m_{[max,avg]}$, resulting in refined features: $O_{[max,avg]} \in \mathbb{R}^{H\times W\times C}$. This process can be described by the following equation:

$$\begin{aligned} m_{[max,avg]} &= \sigma\big(MLP(\mathcal{P}^s_{[max,avg]}(h))\big) \odot h \\ &= A^c_{[max,avg]} \odot h \end{aligned} \tag{2}$$

$$\begin{aligned} O_{[max,avg]} &= \sigma\big(Conv(\mathcal{P}^c_{[max,avg]}(m))\big) \odot m \\ &= A^s_{[max,avg]} \odot m \end{aligned} \tag{3}$$

where $\mathcal{P}^s_{[max,avg]}$ and $\mathcal{P}^c_{[max,avg]}$ represents spatial dimension pooling operation and channel dimension pooling on the max branch and the average branch respectively. $\odot$ denotes to hadamard product, and σ is the sigmoid function. $Conv$ is 1×1 convolution layer.

Finally, we employ the dynamic calibration to finely adjust the focus of the attention mechanism, enhancing the model's sensitivity and accuracy towards key information. Further the output of O_d is the weighted sum of refined features $O[max, avg]$ from the two branches. as depicted in the equation:

$$O_d = \omega_{max} O_{max} + \omega_{avg} O_{avg} \tag{4}$$

where ω_{avg} and ω_{max} are learnable scalars to control the relative importance of two branches. Subsequently, a 1×1 convolution is applied to compress the channel dimensions, and a sigmoid activation function is utilized to generate the attention maps: $A_d \in \mathbb{R}^{C\times 1\times 1}$. These attention maps is weight the original features h:

$$\begin{aligned} O_{hap} &= ConvBlock(\sigma(c * O_d) \odot h) \\ &= ConvBlock(A_d \odot h) \end{aligned} \tag{5}$$

where $c*$ denotes 1×1 convolution layer; $ConvBlock(\cdot)$ is a convolutional block, which helps to integrate these dependencies.

3.4 Joint Preservation

To effectively aggregate spatial-channel dependencies, we analyze both self-attentional perception and hierarchical-attentional perception, implementing lightweight processing on each. While retaining key features of different dimensions, we conducted complementary fusion of these features.

Enhancing Self-Attentional Perceptron (ES). The self-attention mechanism establishes dependencies across arbitrary positions within features, but it struggles to capture inter-channel correlations. Therefore, implementing global modeling on the channel dimension of the features following the self-attention block can effectively complement this limitation. As shown in Figure 4, we use the similar self-attention mechanism to capture the channel dependencies [12]. Specifically, given the feature map $S_{in} \in \mathbb{R}^{H\times W\times C}$ from the self-attention block, we first reshape S_{in} into two branches, resulting in $S_1 \in \mathbb{R}^{C\times(H\times W)}$ and $S_2 \in \mathbb{R}^{(H\times W)\times C}$. Then, we use matrix multiplication between S_1 and S_2, and apply a softmax layer to generate the channel attention maps: $W_c \in \mathbb{R}^{C\times C}$ as:

$$w_{i,j} = \frac{\exp(S_{1,i} \cdot S_{2,j})}{\sum_{i=1}^{C} \exp(S_{1,i} \cdot S_{2,j})} \tag{6}$$

where w_{ij} denotes the impact of i^{th} channel on j^{th} channel. Subsequently, we reshape original feature S_{in} to $S_3 \in \mathbb{R}^{C\times(H\times W)}$, and perform a matrix multiplication between S_3 and W_c. Finally, we apply an element-wise addition operation between S_{in} and the result of the matrix multiplication to obtain the output $es \in \mathbb{R}^{H\times W\times C}$:

$$es_j = \lambda \sum_{i=1}^{C} w_{i,j} S_{3,j} + S_{in,j} \tag{7}$$

where λ is a scale parameter that modulates the significance of the channel attention map in relation to the input feature map, and it progressively learns a weight from 0.

Enhancing Hierarchical-Attentional Perceptron (EH). The hierarchical-attention emphasizes the dependencies within channel and spatial dimensions in localized regions, yet it overlooks interactions on a larger scale. Therefore, incorporating non-local interactions on top of extracting local fine features can enhance the model's ability to accurately segment the target. Inspired by previous works [40, 6], we introduce the EH module, designed to augment non-local interactions in the hierarchical attention block, as illustrated in Figure 5. We first aggregate feature information though Context Modeling, which leverages the inherent correlations within the input feature $H_{in} \in \mathbb{R}^{H\times W\times C}$, deriving global correlation maps: $A_g \in \mathbb{R}^{C\times 1\times 1}$:

$$A_g = \psi\big((\mathcal{F}_{r1}(c_1 * H_{in})\big) \times \mathcal{F}_{r2}(H_{in}) \tag{8}$$

where $\psi(\cdot)$ is softmax function; c_1* is a convolution layer with 1×1 kernel size; $\times$ denotes matrix multiplication and $\mathcal{F}_{ri}(\cdot)$ denotes reshaping operations. Subsequently, instead of the previous methods that distribute the global correlation map across all pixel positions via element-wise addition, we introduce Context Integration for global interactions. This method transforms the global correlation maps A_g into global attention maps $W_g \in \mathbb{R}^{C\times 1\times 1}$. These maps are then used to weight the original feature H_{in}, effectively integrating non-local information into the feature representation:

$$\begin{aligned} eh &= \sigma\Big(c_2 * GELU\big(LN(c_1 * A_g)\big)\Big) \odot H_{in} \\ &= W_g \odot H_{in} \end{aligned} \tag{9}$$

Table 1. Results on Synapse multi-organ dataset. DICE scores are reported for individual organs. ↑ denotes higher the better, ↓ denotes lower the better. The best results are in bold.

Method	Average			GFLOPs	Aorta	GB	KL	KR	Liver	PC	SP	SM
	DICE↑	mIoU↑	ASD↓									
UNet[22]	70.11	59.39	14.41	65.41	84.00	56.70	72.41	62.64	86.98	48.73	81.48	67.96
AttnUNet [20]	71.70	61.38	10.00	67.11	82.61	61.94	76.07	70.42	87.54	46.70	80.67	67.66
TransUNet [7]	77.48	67.32	4.66	28.39	87.23	63.13	81.87	77.02	94.08	55.86	85.08	75.62
SSFormer [33]	78.01	67.23	4.56	29.71	82.78	63.74	80.72	78.11	93.53	61.53	87.07	76.61
MT-UNet [32]	78.59	66.33	5.10	44.21	87.92	64.99	81.47	77.29	93.06	59.46	87.75	76.81
LeViT-UNet[36]	78.53	67.17	4.40	25.55	87.33	62.23	84.61	80.25	93.11	59.07	88.86	72.76
SwinUNet [5]	79.13	66.88	4.70	26.25	85.47	66.53	83.28	79.61	94.29	56.58	90.66	76.60
CASTformer [37]	82.55	74.69	5.81	42.71	**89.05**	67.48	86.05	82.17	**95.61**	67.49	91.00	81.55
CASCADE [21]	82.68	73.48	2.83	44.83	86.63	68.48	87.66	84.56	94.43	65.33	90.79	83.52
DANIE-S (our)	84.21	75.42	3.65	13.29	87.31	73.11	85.29	83.86	95.20	70.20	92.85	85.90
DANIE-L (our)	**85.47**	**76.60**	**2.76**	26.55	88.17	**77.80**	**88.59**	**84.68**	94.84	**71.94**	**91.95**	**85.82**

Table 2. Results on the ACDC dataset. DICE scores are reported for individual organs. The best results are in bold.

Method	mDICE ↑	RV↑	Myo↑	LV ↑
UNet[22]	89.41	87.77	85.88	94.67
AttnUnet[23]	89.01	87.30	85.07	94.66
PraNet [11]	90.19	87.21	88.73	94.54
TransUNet [7]	89.71	88.86	84.53	95.73
nnUnet [16]	91.61	90.24	89.24	95.65
MT-UNet [32]	90.43	86.64	89.04	95.62
SwinUNet [5]	90.00	88.55	85.62	95.83
LeViT-UNet[36]	90.32	89.55	87.64	93.76
CASCADE [21]	91.63	89.14	90.25	95.50
DANIE-S (our)	91.96	89.61	90.37	95.89
DANIE-L (our)	**92.36**	**90.12**	**90.99**	**96.00**

where $\sigma(\cdot)$ is sigmoid function; $LN(\cdot)$ is layer normalization; c_i* denotes 1×1 convolution layer. Essentially, EH aims to learn and understand correlations across global regions by aggregating local contextual information provided by the hierarchical attention block.

Aggregation. Finally, we fuse the features from two distinct perceptron modules as follows:

$$e_i = es_i + eh_i \tag{10}$$

these fused features, denoted as: $e_i \in \mathbb{R}^{H\times W\times C}$, are then passed into a convolutional block along with features $j_{i-1} \in \mathbb{R}^{H\times W\times C}$ from the previous Joint Preservation block:

$$j_i = ConvBlock\big(Cat[e_i, j_{i-1}]\big) \tag{11}$$

where $Cat[\cdot,\cdot]$ represents the concatenation of channel dimensions. $ConvBlock(\cdot)$ is a convolutional block, which consists of two 3×3 convolution layers each followed by a batch normalization layer and a GELU activation layer.

We generate four prediction maps from the four stages of the Joint Preservation. The final prediction map, denoted as: y, is computed using additive aggregation:

$$y = \alpha_1 p_1 + \alpha_2 p_2 + \alpha_3 p_3 + \alpha_4 p_4 \tag{12}$$

where p_1, p_2, p_3, and p_4 represent the feature maps from the four prediction heads, and α_i are the weights for individual prediction heads. In our experiments, we set all α values to 1.0.

4 Experiment

4.1 Datasets

Synapse Multi-Organ Segmentation. There are 30 abdominal CT scans with 3779 axial contrast-enhanced abdominal CT images in the Synapse multi-organ dataset [17]. Each CT scan consists of 85-198 slices of 512×512 pixels, with a voxel spatial resolution of ([0:54-0:54]×[0:98-0:98]×[2:5-5:0])mm^3. Following TransUNet [7], we partitioned the dataset into two sets: 18 scans (consisting of 2212 axial slices) for training and 12 scans for validation.

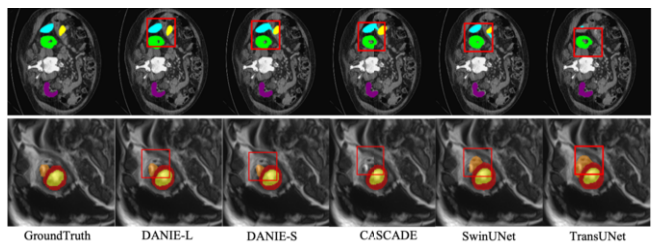

Figure 6. Visualization of segmentation from DANIE and other SOTA methods on Synapse (above) and ACDC (below) dataset. Red rectangles highlight areas where our proposed method clearly outperforms others.

ACDC dataset. The ACDC dataset [3] contains MRI images of 100 patients. The labels for each case include the left ventricle (LV), right ventricle (RV), and myocardium (Myo). Following TransUNet [7], the dataset is split into 70 for training, 10 validation for validation, and 20 cases for testing.

Skin Lesion Segmentation. We conduct extensive experiments on the skin lesion segmentation datasets. Specifically, we utilize the ISIC 2017 dataset [9] comprising 2000 dermoscopic images for training, 150 for validation, and 600 for testing. Moreover, we adopt the ISIC 2018 [8] and follow the literature work [1] to divide the dataset into the train, validation, and test sets accordingly.

GlaS dataset. The GlaS dataset [25] comprises microscopic images of slides stained with Hematoxylin and Eosin (H&E). The dataset includes a total of 165 images, divided into two sets: 85 images designated for training purposes and 80 images allocated for testing.

4.2 Implementation Details

In Synapse multi-organ segmentation, we train each model with a maximum of 300 epochs and a batch size of 24. In ACDC cardiac organ segmentation, we set the batch size to 12 and the maximum number of training epochs to 400 for each model. In ISIC 2017 and 2018, we train our model for 200 epochs with a batch size of 8. In GlaS dateset, we train DANIE with a 100 epochs.

DANIE-S utilizes multi-head attention from MaxViT-tiny [31] as its self-attentional stream, with channel dimensions set to [64, 128, 256, 512] across its encoder layers. Similarly, DANIE-L uses

Table 3. Performance comparison of the proposed method against the SOTA approaches on ISIC 2017 and ISIC 2018 skin lesion segmentation benchmarks. The best results are in bold.

Method	GFLOPs	ISIC 2017				ISIC 2018			
		DICE↑	SE↑	SP↑	ACC↑	DICE↑	SE↑	SP↑	ACC↑
UNet [22]	65.41	81.59	81.72	96.80	91.64	85.45	88.00	96.97	94.04
AttnUNet [20]	67.11	80.82	79.98	97.76	91.45	85.66	86.74	**98.63**	93.76
DAGAN [18]	62.12	84.25	83.63	97.16	93.04	88.07	90.72	95.88	93.24
TransUNet [7]	28.39	81.23	82.63	95.77	92.07	84.99	85.78	96.53	94.52
FAT-Net [35]	23.06	85.00	83.92	97.25	93.26	89.03	91.00	96.99	95.78
TransFuse[38]	26.21	87.31	84.22	96.13	93.50	88.27	90.76	96.11	95.23
TMU-Net [2]	30.49	91.64	91.28	97.89	96.60	90.59	90.38	97.46	96.03
SwinUNet [5]	26.25	91.83	91.42	97.98	97.01	89.46	90.56	97.98	96.15
HiFormer [14]	19.21	92.53	91.55	98.40	97.02	91.02	91.19	97.55	96.21
DANIE-S (our)	13.29	93.84	92.92	**98.73**	97.58	93.32	**95.34**	96.63	96.28
DANIE-L (our)	26.55	**94.25**	**94.10**	98.61	**97.72**	**93.47**	94.17	97.27	**96.43**

Table 4. Performance comparison of the proposed method against the SOTA approaches on GlaS dataset. The best results are in bold.

Method	GFLOPs	mDICE↑	mIoU↑
UNet [22]	65.63	89.62	82.14
UNet++ [39]	138.66	90.05	82.80
AttnUnet [20]	66.64	89.90	82.87
TransUNet [7]	28.39	91.08	84.01
TransFuse[38]	26.21	90.97	84.29
SSFormer[33]	29.31	91.08	83.75
SwinUNet [5]	26.25	91.39	84.60
TGANet[26]	42.08	91.43	84.90
CASCADE [21]	44.83	90.97	84.76
DANIE-S (our)	13.29	91.28	84.23
DANIE-L (our)	26.55	**92.65**	**86.28**

MaxViT-small model for its self-attention stream, featuring channel dimensions of [96, 192, 384, 768] in the encoder. Both models initialize using pre-trained ImageNet weights from the timm library for MaxViT. We train our model using AdamW optimizer with a weight decay and learning rate of 0.0001. We optimize the combined DICE and Cross-Entropy (CE) loss $\mathcal{L}$ with $\lambda_1 = 0.7$ and $\lambda_2 = (1\text{-}\lambda_1) = 0.3$ in all our experiments:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{dice} + \lambda_2 \mathcal{L}_{ce} \tag{13}$$

where λ_1 and λ_2 are the weight for the DICE ($\mathcal{L}_{dice}$) and CE ($\mathcal{L}_{ce}$) losses, respectively. Our framework was implemented with Pytorch and all experiments were performed on an NVIDIA GeForce RTX 3090 GPU.

4.3 Evaluation Metrics

We use mDICE, mean intersection over union (mIoU) and Average surface distance (ASD) as the evaluation metrics in our experiments on Synapse Multi-organ dataset. Following existing methods, we use DICE scores for the ACDC dataset. In the experiment of the Skin Lesion Segmentation, we made an overall evaluation of the mainstream medical image segmentation network by using four indicators: mDice, Sensitivity (SE), Accuracy (ACC), and Specificity (SP). In GlaS dataset, we apply mDICE and mIoU as the evaluation metric in our experiments.

4.4 Comparative Results

Comparative Results on Synapse dataset. Table 1 shows the performance comparison of our DANIE with state-of-the-art methods

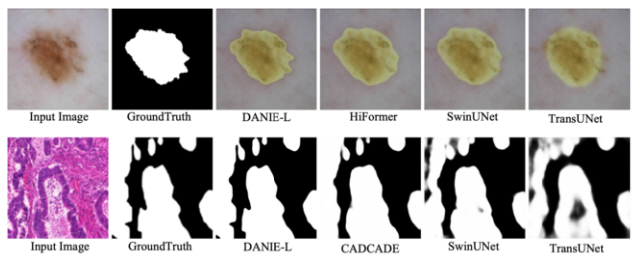

Figure 7. Visual comparisons of different methods on the ISIC 2017 (above) and GlaS (below) dataset.

on the Synapse dataset. Specifically, compared to the mainstream segmentation method TransUNet, our DANIE-L achieved improvements of 7.99%, 9.28%, and 1.9% in average DICE, mIoU, and ASD scores. It also can be observed that DANIE-L outperforms the recent best model CASCADE [21], with absolute improvements of 2.79% and 3.12% in average Dice and mIoU respectively. Our DANIE model achieved higher scores than the SOTA model on 6 out of the 8 organs, which is sufficient to demonstrate that DANIE has adequate performance to accurately segment both large and small organs. Additionally, compared to previous transformer-based methods, DANIE is more advantageous in segmentation gallbladder, pancreas and stomach, which are difficult to delineate using past segmentation models. It can be observed that DANIE not only accurately localizes organs but also produces coherent boundaries, even in small object.

Comparative Results on ACDC dataset. From the results shown in Table 2. DANIE-L outperforms two popular methods, TransUNet and SwinUNet, by 2.65% and 2.36% respectively. Additionally, DANIE-L gains average DICE score (92.36%), RV DICE score (90.17%), Myo DICE score (90.99%), and LV DICE score (96.00%) are all superior to other SOTA methods. It can be seen that our method also demonstrates strong performance on MRI images of the human heart, verifying our promising scalability across different medical imaging data modalities.

Comparative Results on Skin Lesion Segmentation datasets. The comparison results for benchmarks of ISIC 2017 and ISIC 2018 skin lesion segmentation task against leading methods is presented in Table 3. In ISIC 2017 dataset, we can observe that our DANIE-L achieve the highest average DICE (94.25%), SE (94.10%) and ACC (97.72%) surpassing the current SOTA method HiFormer [14] by 1.72%, 2.55% and 0.7%. In ISIC 2018 dataset, the mDice value of our DANIE-L is 2.45%, 4.01% and 8.48% higher than that of the HiFormer, SwinUNet and TransUNet network respectively.

Table 5. Ablation study of the Dual Attention Encoder on the Synapse dataset. spatial/channel att denotes spatial/channel attention from hierarchical-attention. The best result is bolded.

Index	Architecture	GFLOPs	Avg DICE
A.1	DANIE w/o hierarchical-attention	19.06	77.12
A.2	DANIE w/o self-attention	10.98	74.06
A.3	DANIE w/o spatial att	22.13	81.18
A.4	DANIE w/o channel att	23.25	80.65
A.5	DANIE	25.16	**82.97**

Comparative Results on GlaS dataset. We conducted experiments on GlaS dataset, which focus on segmenting glands from stained slide images. As shown in Table 4, our DANIE still outperforms other competitive methods. Specifically, DANIE-L get highest scores in mDice (92.65%) and mIoU (86.28%) than other SOTA methods. Note that TransFuse adopts attention mechanism in decoder for capturing multi-scale information, which ignores the complementarity between dependencies across different dimensions, and thus obtain sub-optimal performance.

4.5 Visualization Results

We visualize some segmentation results of our DANIE and other methods on four public datasets. As shown in Figure 6, our DANIE can greatly reduce the number of false positive predictions than TransUNet [7] on the Synapse dataset. Especially for the prediction of red block, our results are more consistent with the original image, and there are fewer areas of wrong prediction. Meanwhile, compared to SwinUNet [5] and CASCADE [21] on ACDC dataset, we can see that DANIE produced more accurate prediction maps for the right ventricle. In fact, this phenomenon has been reflected in Table 1 and Table 2, where DANIE is significantly better than CASCADE when DICE is the default evaluation metric. For the large lesion regions segmentation, such as ISIC 2017 and GlaS datasets as shown in Figure 7. It shows that our results are more distinguishable than other networks. DANIE generates better segmentation results, which are more similar to the ground truth than other SOTA methods. These results verify that our model works well with different types images.

4.6 Ablation Studies

Effectiveness of Dual Attention Encoder. We conducted a series of ablation studies with various structures to validate the effectiveness of the Dual Attention Encoder, To ensure the fairness, we simply used element-wise addition for fusion.

As shown in Table 5, we observe a notable improvement in performance with an increase in the dimensionality of feature extraction. Through A.1, A.3, and A.4, we can see that adding local attention mechanisms for additional dimensions to a self-attentional stream significantly boosts performance. This finding confirms that extracting dependencies from both global and local perspectives aids the model in more effectively segmenting details in images. Additionally, gradual enhancement in segmentation accuracy with the expansion of attention dimensions within the model. This suggests that, unlike mainstream methods that primarily rely on self-attention for capturing feature dependencies in a single dimension, integrating multiple dependencies proves to be more effective.

Effectiveness of Hierarchical-Attentional Perception. In this experiment, we employed the same experimental setup as in the Dual Attention Encoder ablation study. We analyzed the impact of component choreography order in hierarchical-attention on the model, since the different functions of each module mean order may affect overall performance. Table 6 summarizes the experimental results of different attention order and dynamic calibration. From the results, channel-first order performs slightly better than spatial-first order. Meanwhile, we found that when dynamic calibration is included in hierarchical attention, it can better refine features, thus improving segmentation accuracy. This indicates the advantages of hierarchical attention in the order design of channel and spatial attention as well as dynamic calibration.

Table 6. Ablation study of the Hierarchical-Attentional Perception structure. Dcal: dynamic calibration. The best result is bolded.

Index	Description	Dcal	Synapse	ACDC
B.1	Spatial + Channel	$\times$	81.52	90.43
B.2	Spatial + Channel	mc	82.22	90.79
B.3	Channel + Spatial	$\times$	82.10	90.83
B.4	Channel + Spatial	✓	**82.97**	**91.02**

Table 7. Ablation study of the Joint Preservation on the Synapse dataset. Add: Element-wise addition. Res: Residual.

Index	Fusion	GFLOPs	Avg DICE
C.1	Concat+Res	27.88	83.58
C.2	Add	25.16	82.97
C.3	EH + Add	26.53	84.56
C.4	ES + Add	26.52	84.25
C.5	Joint Preservation	26.55	85.47

Effectiveness of Joint Preservation. Table 7 clearly demonstrates the effects of different fusion mechanisms. All the experiments were conducted under the same backbone and settings. By comparing C.5 with C.1 and C.2, we can see that Joint Preservation outperforms the two current mainstream fusion mechanisms (element-wise addition and concatenation). Despite a slight increase in complexity over C.2, the average DICE on the Synapse dataset improved by 2.5%. We believe the extra computation is worthwhile. Furthermore, comparing C.2 with C.3 and C.4, we can assert that Joint Preservation effectively enhances the fusion of dual attentional streams, both in lightweight processing of self-attentional and hierarchical-attentional perception.

5 Conclusions

In this work, we introduce DANIE, a simple yet powerful network for medical image segmentation. The key insight is to harness the synergy of various dependencies for effectively modeling semantic information. DANIE benefits from our proposed dual attention encoder that progressively and selectively learns interesting parts of the objects. Further, a joint preservation design is used to boost segmentation performance and correspondingly enable the encoder to capture complementary features. Extensive experiments demonstrate that our DANIE outperforms previous state-of-the-art methods on five popular medical datasets considerably, achieving an optimal balance between computational complexity and segmentation accuracy.

Acknowledgements

This work was supported in part by the National Key Research and Development Program of China under Grant 2020YFA0714003, in part by the Science and Technology Planning Project of Sichuan Province under Grant 2021YFQ0059, and in part by the National Natural Science Foundation of China under Grant No. 61471250.

References

[1] R. Azad, M. Asadi-Aghbolaghi, M. Fathy, and S. Escalera. Bi-directional convlstm u-net with densley connected convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pages 0–0, 2019.

[2] R. Azad, M. Heidari, Y. Wu, and D. Merhof. Contextual attention network: Transformer meets u-net. In *International Workshop on Machine Learning in Medical Imaging*, pages 377–386. Springer, 2022.

[3] O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, X. Yang, P.-A. Heng, I. Cetin, K. Lekadir, O. Camara, M. A. G. Ballester, et al. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE transactions on medical imaging*, 37(11):2514–2525, 2018.

[4] A. Buades, B. Coll, and J.-M. Morel. A non-local algorithm for image denoising. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 2, pages 60–65. Ieee, 2005.

[5] 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 *ECCV*, pages 205–218. Springer, 2022.

[6] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pages 0–0, 2019.

[7] 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.

[8] N. Codella, V. Rotemberg, P. Tschandl, M. E. Celebi, S. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*, 2019.

[9] N. C. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *ISBI 2018*, pages 168–172. IEEE, 2018.

[10] B. Dong, W. Wang, D.-P. Fan, J. Li, H. Fu, and L. Shao. Polyp-pvt: Polyp segmentation with pyramid vision transformers. *arXiv preprint arXiv:2108.06932*, 2021.

[11] D.-P. Fan, G.-P. Ji, T. Zhou, G. Chen, H. Fu, J. Shen, and L. Shao. Pranet: Parallel reverse attention network for polyp segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 263–273. Springer, 2020.

[12] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[13] Y. Han, X. Li, B. Wang, and L. Wang. Boundary loss-based 2.5 d fully convolutional neural networks approach for segmentation: a case study of the liver and tumor on computed tomography. *Algorithms*, 14(5):144, 2021.

[14] M. Heidari, A. Kazerouni, M. Soltany, R. Azad, E. K. Aghdam, J. Cohen-Adad, and D. Merhof. Hiformer: Hierarchical multi-scale representations using transformers for medical image segmentation. In *WACV*, pages 6202–6212, 2023.

[15] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[16] 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*, 18(2):203–211, 2021.

[17] B. Landman, Z. Xu, J. Igelsias, M. Styner, T. Langerak, and A. Klein. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In *Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge*, volume 5, page 12, 2015.

[18] B. Lei, Z. Xia, F. Jiang, X. Jiang, Z. Ge, Y. Xu, J. Qin, S. Chen, T. Wang, and S. Wang. Skin lesion segmentation via generative adversarial networks with dual discriminators. *Medical Image Analysis*, 64:101716, 2020.

[19] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[20] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.

[21] M. M. Rahman and R. Marculescu. Medical image segmentation via cascaded attention decoding. In *WACV*, pages 6222–6231, 2023.

[22] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. *arXiv preprint arXiv:1505.04597*, 2015.

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

[24] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[25] K. Sirinukunwattana, J. P. Pluim, H. Chen, X. Qi, P.-A. Heng, Y. B. Guo, L. Y. Wang, B. J. Matuszewski, E. Bruni, U. Sanchez, et al. Gland segmentation in colon histology images: The glas challenge contest. *Medical image analysis*, 35:489–502, 2017.

[26] N. K. Tomar, D. Jha, U. Bagci, and S. Ali. Tganet: Text-guided attention for improved polyp segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 151–160. Springer, 2022.

[27] Y. Tu, L. Li, L. Su, K. Lu, and Q. Huang. Neighborhood contrastive transformer for change captioning. *IEEE Transactions on Multimedia*, 25:9518–9529, 2023.

[28] Y. Tu, L. Li, L. Su, Z.-J. Zha, C. Yan, and Q. Huang. Self-supervised cross-view representation reconstruction for change captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2805–2815, 2023.

[29] Y. Tu, L. Li, L. Su, Z.-J. Zha, and Q. Huang. Smart: Syntax-calibrated multi-aspect relation transformer for change captioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(7):4926–4943, 2024.

[30] Y. Tu, L. Li, L. Su, Z.-J. Zha, C. Yan, and Q. Huang. Context-aware difference distilling for multi-change captioning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7941–7956, 2024.

[31] Z. Tu, H. Talebi, H. Zhang, F. Yang, P. Milanfar, A. Bovik, and Y. Li. Maxvit: Multi-axis vision transformer. In *European conference on computer vision*, pages 459–479. Springer, 2022.

[32] H. Wang, S. Xie, L. Lin, Y. Iwamoto, X.-H. Han, Y.-W. Chen, and R. Tong. Mixed transformer u-net for medical image segmentation. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2390–2394, 2022. doi: 10.1109/ICASSP43922.2022.9746172.

[33] J. Wang, Q. Huang, F. Tang, J. Meng, J. Su, and S. Song. Stepwise feature fusion: Local guides global. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 110–120. Springer, 2022.

[34] X. Wang, R. Girshick, A. Gupta, and K. He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018.

[35] H. Wu, S. Chen, G. Chen, W. Wang, B. Lei, and Z. Wen. Fat-net: Feature adaptive transformers for automated skin lesion segmentation. *Medical image analysis*, 76:102327, 2022.

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

[37] C. You, R. Zhao, F. Liu, S. Dong, S. Chinchali, U. Topcu, L. Staib, and J. Duncan. Class-aware adversarial transformers for medical image segmentation. *Advances in Neural Information Processing Systems*, 35: 29582–29596, 2022.

[38] 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*, pages 14–24. Springer, 2021.

[39] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE transactions on medical imaging*, 39(6):1856–1867, 2019.

[40] Z. Zhu, M. Xu, S. Bai, T. Huang, and X. Bai. Asymmetric non-local neural networks for semantic segmentation. In *ICCV*, pages 593–602, 2019.