Effective and Efficient Medical Image Segmentation with Hierarchical Context Interaction

Zehua Cheng[1], Di Yuan[1], Wenhu Zhang[2], Thomas Lukasiewicz[1]
[1]Department of Computer Science
University of Oxford
[2]FLock.io
zehua.cheng@cs.ox.ac.uk

Abstract

The U-Net models have become the predominant architecture within the domain of medical image segmentation. Recent advancements have showcased the potential of incorporating attention-based techniques into U-Net structures. Nevertheless, the inclusion of attention mechanisms often leads to a substantial increase in both computational demands and the number of parameters, with only a marginal improvement in the performance. This observation raises a critical evaluation of the efficiency associated with the integration of attention modules. In this paper, we propose a novel methodology termed Hierarchical Context Interaction *(HCI), a parameter-efficient, attention-free enhancement that can be seamlessly incorporated into U-Net-based models. Experimental results demonstrate that our proposed HCI module attains state-of-the-art performance on two widely used benchmarks, i.e. Medical Segmentation Decathlon Datasets and Synapse Datasets, while concurrently sustaining a computationally efficient profile comparable to conventional U-Net configurations.*

1. Introduction

Compared to 2D images, 3D images offer a more comprehensive view of the structures of interest, encompassing spatial, volume, morphological, dynamic, and individual information [33]. 3D medical image segmentation is a fundamental task in clinical applications, aiming to partition an image into regions corresponding to different anatomical structures or pathological tissues. U-Net [26] is a seminal architecture for medical image segmentation. Beyond its elegant design and adaptability, U-Net's symmetric expansive path ensures the recovery of fine-grained details that are often lost in traditional convolutional neural networks (CNNs), making it particularly suitable for tasks where spatial context and resolution are paramount. Its modular design also facilitates seamless integration with other neural network components, allowing researchers to build upon its foundational structure and adapt it to a wide range of medical imaging challenges. Due to its simple architectural design and high extensibility, U-Net has become one of the most widely used frameworks for medical image segmentation across various modalities.

Intriguing research has revealed that an attention module, such as self-attention, soft-attention, or spatial attention, can be incorporated at any location within the U-Net architecture [2]. The attention mechanism enables the model to focus on relevant features while ignoring irrelevant ones, thereby enhancing segmentation accuracy [28].

However, despite the promising results achieved by integrating attention modules into the U-Net architecture, much of the existing research has concentrated on performance improvement through the addition of more parameters and computational resources. The incorporation of the attention mechanism typically signifies a substantial augmentation in the computational load, often quantified as a several-fold increase in *floating point operations per seconds* (FLOPs) and memory consumption. Nevertheless, augmenting the number of parameters and computational complexity (measured in FLOPs) by a significant factor does not invariably lead to a substantial enhancement in model performance. For instance, our observations indicate that Attention U-Net [24] necessitates a computational overhead several-fold greater than the traditional U-Net, yet it merely manifests 1% increment in performance (measured in the Dice similarity coefficient). Consequently, we postulate the parameter efficiency of the attention mechanism, particularly when it demands a substantial increase in computational resources compared to the baseline model. Specifically, we question the prudence of such an extensive parameter addition for only incremental gains in performance.

In this work, we argue that the role of the attention module in visual medical tasks can be replaced by convolutional

operations, which is more computationally efficient. We revisit U-Nets and propose a novel U-Net-like structure, termed *hierarchical context interaction* (HCI) U-Net. The HCI U-Net can achieve state-of-the-art 3D medical image segmentation performance with HCI modules that consist solely of convolution operators. Drawing upon the inherent simplicity and modularity of the U-Net architecture, we posit that an effective module, constructed using a series of convolutional neural networks (CNNs), can enhance the performance of U-Net in medical image segmentation tasks. By increasing the parameter count by a mere 18% and the computational complexity, measured in floating point operations per second (FLOP), by 40%, we have achieved the current state-of-the-art results. In contrast, the foundational Attention U-Net increases its parameters by 198.28% and FLOPs by 162.06% compared to the standard U-Net, yet it realizes only a 1% improvement in performance on datasets such as Synapse and Medical Segmentation Decathlon.

Analogous to Swin-UNet [4], which replaces the convolutional block with a Swin Transformer layer, we substitute the traditional convolutional block with a *hierarchical context interaction* (HCI) block. This block captures long-range dependencies while maintaining awareness of local structures. In contrast to attention-based methods, which rely on an element-wise affinity matrix in self-attention, our HCI mechanism utilizes a more coarse-grained visual feature interaction scheme. We contend that utilizing visual feature interaction can achieve performance comparable to the self-attention module, but with a lower computational cost.

The main contributions can be summarized as follows. First, we propose HCI UNet, a U-Net-like structure without attention, which significantly improves the 3D medical image segmentation performance of U-Net. Second, we show that, in contrast to attention, the proposed method is computationally efficient in learning global information. Third, the experimental studies show that the proposed method can achieve state-of-the-art performance on all Medical Segmentation Decathlon Datasets [1] with a similar computational cost as vanilla UNet.

2. Related Work

2.1. Recent Convolutional Architectures

Recent advancements in Vision Transformer (ViT) have paved the way for innovative approaches to enhance the performance of CNNs. Several studies have proposed the integration of Transformer-style architectures with spatial convolutions. Given that self-attention is capable of capturing global information, these studies aim to replicate self-attention with spatial convolutions by incorporating large convolutional kernels within a convolutional architecture to learn long-range relations. For instance, ConvNeXt [22] further provides a comprehensive analysis of the design principles inherent in recent vision Transformers, leading to the development of a robust convolutional model that employs 7×7 depth-wise convolutions. Another approach is to introduce explicit high-order spatial interactions into existing architecture [19]. FocalNet [31] constructs an interaction between a sequential gate aggregation and queries that could extract local-global interaction representation. These successful applications expose the potential of a combination of convolution that could outperform self-attention.

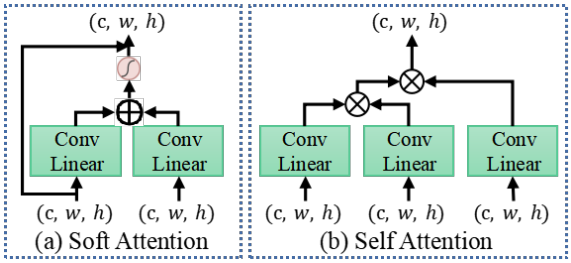


Figure 1. Illustration of attention mechanism.

U-Net structure incorporated a skip-connection to bridge the reference between global and local features which eliminates the need to introduce a larger convolution kernel. We then applied a combination of convolution to explore the potential of bridging explicit high-order spatial interaction.

2.2. Attention-based U-Nets.

U-Net [26] contains a simple U-shaped symmetric structure and skip connections to concatenate the deep and coarse features in the expansive path with the shallow and fine features in the contracting path, it can significantly improve the accuracy of medical image segmentation. In recent years, an increasing number of researchers have focused on introducing self-attention or incorporating Transformer architecture into U-Nets to enhance network performance [29]. Attention U-Net [24] introduces soft attention to the feature maps present within the skip connections. MA-Net [11] [25] employs self-attention for the extraction of local features in relation to global dependencies. [12] has demonstrated the successful combination of spatial pyramid pooling [14] and soft attention to utilize multi-scale features. Swin-UNet [4] uses the Swin Transformer [21] block as the basic unit to construct a U-shaped encoder-decoder architecture with skip connections for medical image segmentation. In conclusion, the introduction of self-attention has achieved significant improvements in the U-Net structure. However, in this work, we replace the self-attention module with our HCI block to enhance representation.

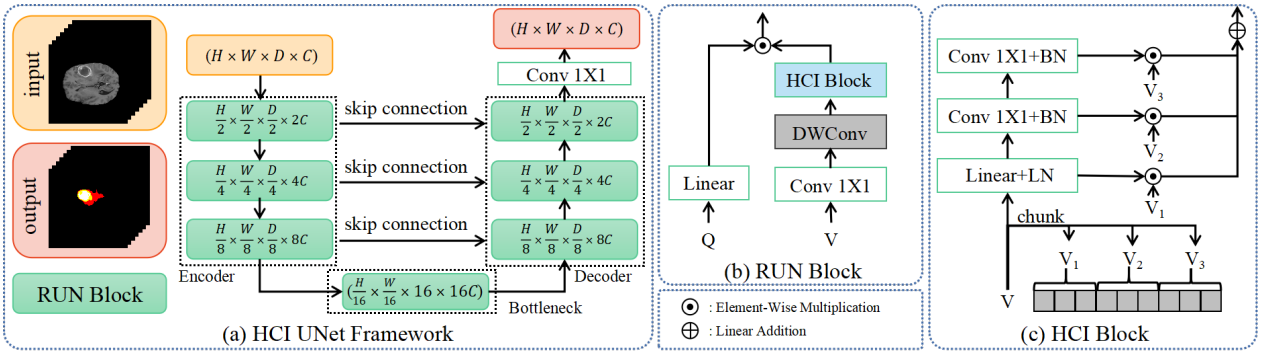


Figure 2. (a) Overview of the HCI UNet, which is similar to Swin-UNet [4], but replaces Transformer blocks with RUN blocks, and does not require additional patch merging layers and patch expanding layers. (b) Our Revisiting U-Net (RUN) block without any attention can directly use Q and V to learn feature representations. (c) Our Hierarchical Context Interaction (HCI) block can extract important information using the multi-sequence method. BN: Batch Normalization [15], LN: Linear Normalization [3], and DWConv: Depthwise Separable Convolution [7].

3. Preliminaries

In the sequel, let $X \in \mathcal{R}^{H\times W\times D\times C}$ be a given input image, where H and W are its height and width, respectively, D is the number of channels (usually, $D > 3$ for 3D medical images), and C is the number of channels for the medical images.

To contrast the differences between our contributions and the traditional attention module, we first introduce the concepts of soft attention and self-attention, shown in Fig. 1.

3.1. Soft Attention

Here, we discuss the attention operation of the original Attention U-Net [24], where for a given layer l and the corresponding feature map of the l-th layer x^l. Formally:

$$\begin{aligned} q_{att}^l &= \psi^T\left(\sigma\left(W_x^T x_i^l + W_g^T g_i + b_g\right)\right) + b_\psi \\ \alpha_i^l &= sigmoid\left(q_{att}^l\left(x_i^l, g_i\right)\right), \end{aligned} \tag{1}$$

where $\sigma\left(x_i^l\right) = \max\left(0, x_i^l\right)$, $sigmoid\left(x_i\right) = 1/(1+\exp\left(-x_i\right))$, $g_i \in \mathcal{R}$ is the gating vector for each pixel i to determine focus regions, W_x^T and W_g^T are the learnable weights for a linear transformations, and b_ψ and b_g are the bias terms.

3.2. Self-Attention

We follow the notations of self-attention [27]. Given a sequence of input tokens, the self-attention mechanism first computes the query matrix $\mathbf{Q}$, key matrix $\mathbf{K}$, and value matrix $\mathbf{V}$ as follows:

$$Q = x_s^l W_q,\ K = x_s^l W_k,\ V = x_s^l W_v, \tag{2}$$

where s_s^l is the input sequence of the image X at the l-th layer, and W_q, W_k, and W_v are learned weight matrices. In a vision Transformer (ViT) [10], usually $\mathbf{Q}$ and $\mathbf{K}$ are based on the same input with different linear projections. The self-attention mechanism then computes an attention score between each pair of tokens in the sequence by taking the dot product between the corresponding query and key vectors:

$$Attention(Q, K, V) = softmax(\frac{Q\cdot K^T}{\sqrt{d_k}})V, \tag{3}$$

where $\sqrt{d_k}$ is to prevent the attention score from becoming too small or large. Self-attention benefits from the element-wise affinity matrix calculated from $\mathbf{Q}\cdot\mathbf{K^T}$. The size of the affinity matrix is proportional to the length of the input, resulting in an $n \times n$ matrix for an input length of n. Consequently, as the input length increases, the computational costs of self-attention also increase substantially.

4. Method

To learn input-dependent long-range interactions, we propose HCI U-Net, a novel U-Net-like framework for medical image segmentation. The structure is presented in Fig. 2. HCI U-Net achieves efficient medical image segmentation without any attention by employing the hierarchical context interaction (HCI) block and revisiting the U-Net (RUN) block.

4.1. Motivation

Self-attention is well-suited for processing long sequences and natural language processing tasks, as it can adaptively learn the correlations between different parts of the input sequence, thereby enhancing the model's performance and generalization ability. We observe that the dot product operations in self-attention can generate a large affinity matrix, leading to a non-trivial memory footprint

and computational cost. Therefore, instead of using the self-attention mechanism, we propose a novel visual feature interaction mechanism. Specifically, we suggest replacing the element-wise affinity matrix in self-attention with a visual feature interaction module. In previous studies, the element-wise affinity matrix has been shown to be an effective solution for extracting fine-grained features. Notably, self-attention performs a dot product between the query matrix $\mathbf{Q}$ and the key matrix $\mathbf{K}$, where in most cases, $\mathbf{Q}$ and $\mathbf{K}$ have the same input with different linear layers. Thus, we hypothesize that using $\mathbf{Q}$ alone can be efficient.

Besides, self-attention uses an element-wise operation to obtain an affinity matrix to realize fine-grained feature interaction. However, recent findings [17] demonstrate that these obtained affinity matrices are very sparse, and maintaining huge affinity matrices is inefficient and unnecessary. Therefore, we propose a hierarchical context interaction (HCI) block to achieve intensive and efficient information interaction. Intuitively, we divide $\mathbf{V}$ into chunks of equal size and perform element-wise multiplication between each block and $\mathcal{F}(Norm(\mathbf{V}))$, where $\mathcal{F}$ can be convolutional layer with a kernel size 1×1 (Conv 1×1) or a linear projection layer, and $Norm$ can be layer normalization [3] or batch normalization [15]. In this way, the interaction of local and global information can be directly used to capture long-range dependencies with a good awareness of local structures.

4.2. Hierarchical Context Interaction

Similar to Swin-UNet [4], we adopt a symmetric structure consisting of an encoder, a decoder, and skip connections. However, different from the Swin Transformer block of Swin-UNet, the proposed revisiting U-Net (RUN) block is constructed based on the hierarchical context interaction (HCI) block. The overall framework is presented in Fig. 2 (a). Fig. 2 (c) and (d) show our RUN and HCI block in detail. Each RUN block takes $\mathbf{Q}$ and $\mathbf{V}$ as inputs. $\mathbf{Q}$ is passed by a linear projection layer. $\mathbf{V}$ is passed through Conv1×1, Depthwise Separable Convolution (DW-Conv) [7], and an HCI block sequentially. Note, we do not require $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ for multi-head attention in the Swin Transformer block, shown in Fig. 2 (b).

In the HCI block, we first obtain $\mathbf{V^1}$, $\mathbf{V^2}$, and $\mathbf{V^3}$ via Eq. (4).

$$\begin{aligned} V^1 &= LN(\mathcal{F}(V)), \\ V^2 &= BN(Conv(V^1)), \\ V^3 &= BN(Conv(V^2)), \end{aligned} \tag{4}$$

where $\mathcal{F}$ is the linear projection, LN is the layer normalization operation [3], $Conv$ is the 1×1 convlution [20], BN is the batch normalization operation [15], and $\odot$ is the element-wise multiplication. We then obtain $\mathbf{V_1}$, $\mathbf{V_2}$, and $\mathbf{V_3}$ through the chunk operation on $\mathbf{V}$. ($\mathbf{V^1}$, $\mathbf{V^2}$, $\mathbf{V^3}$) and ($\mathbf{V_1}$, $\mathbf{V_2}$, $\mathbf{V_3}$) are multiplied by dot product, respectively, to get $\tilde{V}$, which learns the hierarchical context interaction information. Based on the HCI block, $\tilde{\mathbf{V}}$ can be expressed as:

$$\tilde{V} = V^1 \odot V_1 + V^2 \odot V_2 + V^3 \odot V_3, \tag{5}$$

Finally, the output of the RUN block is then the element-wise multiplication of $\mathbf{Q}$ and $\tilde{\mathbf{V}}$:

$$Out = \mathcal{F}(Q) \odot \tilde{V}. \tag{6}$$

4.3. Network Architecture

In the encoder, the inputs with the resolution of $H \times W \times D \times C$ are fed into the RUN block. We use a DW-Conv and Conv 1×1 to reduce the number of tokens ($2\times$ downsampling) and increase the feature dimension to $2\times$. This operation is repeated three times in the encoder. In the bottleneck (see the bottom of Fig.2 (a)), we only use one RUN block to learn the representations. For the bottleneck, the input and output dimensions remain unchanged. For the decoder, which is symmetrical to the encoder, we use DW-Conv and Conv 1×1 to reshape the higher-resolution feature map ($2\times$ upsampling) and reduce the feature dimensions by half, accordingly.

Apart from the HCI block and RUN block, the proposed method replaces the pooling layers in the encoder of U-Net with Conv 1×1. Or, similarly, HCI UNet replaces the patch merging layers and patch expanding layers in the encoder and decoder of Swin-UNet with Conv 1×1. Compared with the pooling layers in U-Net, Conv 1×1 can learn more meaningful representations; compared with the patch merging layers and patch expanding layers in Swin-UNet, Conv 1×1 is computationally more efficient in GPU processing than pooling operation or linear projection [6]. The skip connections in UNet are inherited to fuse the multi-scale features from the encoder with the upsampled features to reduce the loss of spatial information caused by down-sampling.

5. Experiments

5.1. Datasets

We conduct the experiments on Medical Segmentation Decathlon Datas-et (MSD) [1], which is a benchmarking platform for evaluating the performance of medical image segmentation algorithms. All baselines are evaluated in all MSD challenge datasets, including Brain, Hippocampus, Liver, Lung, heart, Pancreas, Hepatic Vessel, Spleen, Prostate, and Colon. Brain and Prostate are mp-MRI datasets, Hippocampus and Heart are MRI datasets, and other datasets are CT datasets. Besides, we also conduct the experiments on Synapse Multi-Organ Segmenta-

tion Dataset (Synapse) [1], which serves as a valuable resource for training and evaluating multi-organ segmentation models with a large-scale collection of images and corresponding organ annotations. It includes annotations for 13 organs, i.e., heart, liver, and lungs, and encompasses diverse medical image modalities, including CT and MRI scans. For each dataset, we randomly split the data into training (80%) and testing (20%) sets. In particular, we use the Dice similarity coefficient (DSC) and 95% Hausdorff Distance (HD95) on the MSD Brain Dataset [4] and use DSC and Normalized Surface Distance (NSD) on the CT datasets [32]. The higher the values of DSC and NSD are, the smaller the value of HD is.

5.2. Implementation

If not specified, all models are implemented on four NVIDIA A100 GPUs. We use Dice loss [23] as the loss function and train all models with Adam optimizer [18] with a learning rate of 10^{-3} and the weight decay as 10^{-5}. If not specified, the batch size is 2 per GPU. We use the original implementation [2] and follow the official training schedule of the nnUNet [16] [30]. We only report the best performance of nnUNet. For Swin-UNet, we use the default training strategy of the original paper. The input resolution of Swin-UNet is 224×224. For all other models, follow the official implementations in MONAI [3]. We apply $1,000$ epochs to train the UNETR with an initial learning rate 10^{-4}. Except for Swin-UNet [4], which uses weights pre-trained from ImageNet [9], all experiments are trained from scratch.

It is worth noting that we follow the official model configurations of nnUNet, such as deep supervision and auxiliary loss functions, which is a different setup compared to other baselines. What's more, nnUNet has a unique model architecture for each dataset.

5.2.1 Baselines

To validate that HCI UNet can achieve efficient segmentation performance without relying on attention mechanisms, we compare it with attention-based and attention-free improvement methods. For attention-free methods, we apply UNet3D [8] and nnU-Net [16]. UNet3D is a 3D variant of the original U-Net, while nnU-Net extends the U-Net framework with multiple enhancement modules, adaptive image preprocessing, data augmentation strategies, and automation-driven hyperparameter optimization.

For attention-based methods, we present attention-based U-Net and Trans- former-based U-Net on both datasets, because Transformer is a variant of attention. We present the experiments on the Attention U-Net [24], TransUNet [5], UNETR [13] and Swin-UNet [4].

5.3. Results

5.3.1 Quantitative Comparison on MSD

The results in Table 1 show the overall performance comparison in 10 benchmark MSD Datasets [1]. HCI UNet achieves state-of-the-art performance in both evaluation metrics in all datasets. This suggests that HCI UNet is more effective than attention-based methods and Transformer-based methods. HCI UNet outperforms nnUNet (second highest in mean DSC of the 10 MSD Datasets) by 1.43% and outperforms SwinU-Net (best performing attention-based method) by 2.46%.

By comparing UNet3D [8] and Attention U-Net (AttnUNet) [24], we find that introducing an attention module did not necessarily lead to a significant improvement in performance and, in fact, could result in a decline in performance when dealing with complex objects. These findings suggest that the attention mechanism may not always be effective in 3D medical image segmentation, depending on the specific circumstances. By comparing UNet3D and Transformer-based U-Nets, we observe that TransUNet, which has the least obvious improvement, has even lower performance than UNet3D in the MSD Brain Dataset. While Swin-UNet [4], the best-performing Transformer-based baseline, has a significant improvement, it has much more parameters than UNet3D. These observations suggest that the Transformer-based U-Nets with multi-head attention blocks might not be an optimal solution, too. In specific situations, nnU-Net outperforms Swin-UNet, indicating that attention is not the only method to enhance the performance of U-Net. By summarizing the observations above, we conclude that HCI UNet can obtain robust segmentation performance by analyzing HCI information.

Table 2 summarizes the object-wise segmentation results on 6 multi-object segmentation datasets. HCI UNet can consistently achieve robust segmentation results in both large objects (organs) and small objects (tumors).For all datasets, HCI UNet outperforms state-of-the-art baselines on most objects of interest. These show that hierarchical contextual interactive learning can learn richer information while avoiding information loss during the learning process.

5.3.2 Quantitative Comparison on Synapse

The results in Table 3 show the segmentation performance comparison of 13 organs in Synapse Dataset. HCI UNet achieves state-of-the-art performance on the average evaluation metric, which also shows that HCI UNet is more effective than attention-based methods and Transformer-based methods. For example, HCI UNet outperforms UNTER (second highest in average DSC and the best performing

[1] https://www.synapse.org/#!Synapse:syn3193805/wiki/217789
[2] https://github.com/MIC-DKFZ/nnUNet
[3] https://github.com/Project-MONAI/MONAI

2*Method	Brain		Liver		Lung		Pancreas		Hepatic Vessel		Spleen		Colon		Heart		Hippocampus		Prostate		Avg
	DSC ↑	HD95 ↓	DSC ↑	NSD ↑	DSC ↑	NSD ↑	DSC ↑	NSD ↑	DSC ↑	NSD ↑	DSC ↑	NSD ↑	DSC ↑	NSD ↑	DSC ↑	HD95 ↓	DSC ↑	HD95 ↓	DSC ↑	HD95 ↓	DSC ↑
UNet3D	67.21	9.99	82.11	91.27	63.10	62.51	64.68	81.70	65.41	80.54	90.53	93.52	49.32	62.21	90.72	33.41	84.43	1.89	80.15	23.31	73.77
nnU-Net	69.56	9.65	85.86	94.60	73.97	76.02	67.21	83.81	68.96	82.38	97.40	**99.89**	58.33	68.35	92.77	31.53	**89.66**	**1.28**	82.70	20.59	78.64
AttnUNet	66.51	9.97	82.14	91.83	63.20	63.35	64.90	82.77	65.42	80.67	92.10	94.83	50.80	60.11	90.77	33.44	86.15	1.68	79.65	23.92	74.13
TransUNet	64.42	12.98	82.28	91.79	66.54	66.92	67.39	85.53	68.17	82.05	96.22	97.45	51.05	64.33	91.58	31.92	85.74	1.68	80.60	22.43	75.40
UNETR	71.78	8.76	83.92	93.69	73.54	75.91	67.44	84.62	66.85	81.96	96.22	97.58	55.33	64.92	91.00	31.93	86.58	1.49	80.65	22.84	77.33
Swin-UNet	64.36	11.89	86.63	95.65	73.77	77.00	67.54	85.34	67.71	82.18	96.42	98.88	58.13	66.49	92.13	**21.17**	87.39	1.44	82.02	21.60	77.61
HCI UNet	**78.07**	**7.96**	**86.66**	**95.72**	**77.85**	**81.77**	**68.19**	**85.72**	**69.05**	**83.62**	**97.35**	**99.89**	**60.43**	**70.82**	**92.82**	26.55	87.61	1.55	**82.75**	**20.37**	**80.07**

Table 1. Quantitative comparison on the Medical Segmentation Decathlon Dataset [1]. AttnUNet is an abbreviation for Attention U-Net. The reported number is the mean performance of multiple classes if the dataset contains more than one class. **Bold** denotes the best performance.

3*Method	Brain						Liver				Pancreas				Hepatic Vessel				Hippocampus				Prostate			
	WT		TC		ET		Liver		Tumor		Pancreas		Mass		Vessel		Tumor		Anterior		Posterior		PZ		TZ	
	DSC ↑	HD95 ↓	DSC ↑	HD95 ↓	DSC ↑	HD95 ↓	DSC ↑	NSD ↑	DSC ↑	NSD ↑	DSC ↑	NSD ↑	DSC ↑	NSD ↑	DSC ↑	NSD ↑	DSC ↑	NSD ↑	DSC ↑	HD95 ↓	DSC ↑	HD95 ↓	DSC ↑	HD95 ↓	DSC ↑	HD95 ↓
UNet3D	77.11	9.11	69.93	10.43	54.60	10.44	93.25	95.76	70.96	86.78	78.61	92.85	50.75	70.55	62.32	83.15	68.50	77.92	84.48	1.88	84.38	1.91	72.77	26.50	87.53	20.11
nnU-Net	77.91	10.03	73.33	8.95	57.44	9.97	95.75	98.55	75.97	90.65	**81.64**	**96.14**	52.78	71.47	**66.34**	84.20	71.58	80.55	**90.37**	**1.24**	**88.95**	**1.31**	**75.81**	**23.00**	89.59	18.18
AttnUNet	76.71	9.00	68.51	10.46	54.30	10.45	93.28	96.55	71.00	87.11	78.65	93.15	51.15	72.39	62.33	83.30	68.50	78.03	86.78	1.64	85.52	1.72	72.77	26.34	86.53	21.50
TransUNet	70.63	14.02	68.42	14.50	54.20	10.42	93.45	96.17	71.10	87.41	80.26	95.77	54.51	75.28	65.80	84.00	70.54	80.10	87.05	1.56	84.42	1.80	73.42	24.35	87.77	20.50
UNETR	79.82	8.20	76.51	8.85	59.00	9.22	94.53	98.35	73.30	89.03	81.55	95.47	53.32	73.77	63.34	83.48	70.35	80.44	87.32	1.42	85.83	1.57	73.37	24.90	87.93	20.77
Swin-UNet	70.02	14.20	70.55	10.33	52.50	11.13	**95.70**	**98.40**	77.56	92.90	80.66	96.09	54.41	74.58	64.30	83.77	71.11	80.58	88.33	1.37	86.44	1.51	75.11	23.50	88.93	19.70
HCI UNet	**90.57**	**7.65**	**82.03**	**7.72**	**61.62**	**8.50**	**95.70**	**98.40**	**77.62**	**93.04**	81.57	95.50	**54.81**	**75.93**	65.89	**85.13**	**72.20**	**82.11**	88.10	1.59	87.12	1.51	75.52	23.50	**89.98**	**17.23**

Table 2. Quantitative comparison of segmentation performance on the multi-object datasets [1]. WT, TC, and ET represent the whole tumor region, tumor core region, and tumor enhancement region, respectively. **Bold** denotes the best performance.

Transformer-based method) by 2.48% on average DSC. Then, by comparing UNet3D [8] and Attention U-Net (AttnUNet) [24], we also find that the performance improvement after adding the attention module is not obvious. For example, the average DSC of AttnUNet is only 1.53% higher than that of UNet3D. This means that under certain circumstances (e.g. AttnUNet is 1.13% lower than UNet3D in Veins), the attention mechanism might not be efficient in 3D medical image segmentation. Moreover, by comparing UNet3D and Transformer-based U-Net, we observe that although Transformer-based baselines have significant improvements, they have many more parameters than UNet3D. Compared to the improvement in limited segmentation performance, the increase in the number of parameters may not be worth the candle. These observations suggest that a Transformer-based U-Net with multi-head attention blocks might not be the optimal solution either. In addition, nnU-Net shows better performance than Transformer-based methods under certain circumstances. This also validates our hypothesis that attention is not the only way to improve the performance of U-Net. Finally, we observe that in some datasets (especially the Aorta and pancreas Datasets), HCI UNet segmentation performance is lower than that of the Transformer-based methods. This may be because, for organs with large variations in shape and size and adjacent structural interference (such as blood vessels, intestines, lymph nodes, etc.), the multi-scale attention based on the Transformer method can better handle this variation. Overall, HCI UNet can obtain robust segmentation performance on most datasets by analyzing HCI information.

5.3.3 Qualitative Comparison

We show the qualitative results of Synapse and MSD Brain Dataset [1] in Fig. 3 and Fig. 4, respectively. From top to bottom are the segmentation outputs of three samples of the corresponding datasets, and from left to right are the original inputs, ground truths, the visualization results of baseline models, and our proposed HCI UNet. In Fig. 3, the segmentation performance of HCI UNet is the closest to the ground truth. Especially in TC (green regions), HCI UNet achieves the best segmentation results, while other baselines have obvious under- and/or over-segmentation. Besides, Fig. 4, our method can also achieve relatively ideal segmentation performances for small organs (e.g., the AG) and segmentation objects edges with complex information (e.g., the stomach).

5.3.4 Computational Cost

We analyze the computational composition of baselines and HCI UNet by parameters (params), GPU memory (GPU Mem), and computational cost (GFLOPs). The params refers to the number of adjustable parameters that need to be learned in the model. The GPU Mem refers to the memory space required by the model when using the GPU for training or inference. FLOPs refers to the number of floating-point operands, which is used to measure the calculation amount of the model. The larger the values of these metrics, the higher the computational cost. For all experiments, we calculate the FLOPs and GPU Mem from a single 3D image of a single channel with an input size of $(160, 160, 160)$. We also fixed the depth of the UNet blocks to 5 blocks with in-

Methods	Spl	RKid	LKid	Gall	Eso	Liv	Sto	Aor	IVC	Veins	Pan	AG	DSC Avg
UNet3D	86.70	68.70	77.77	62.53	53.67	93.43	75.58	87.77	64.63	69.19	52.18	60.66	71.07
nnUNet	94.20	89.41	91.02	70.42	72.33	94.76	82.38	87.77	78.21	72.03	68.05	61.63	80.18
AttnUNet	87.30	71.11	77.98	68.88	55.51	93.57	75.75	89.55	66.61	68.06	55.51	61.33	72.60
TransUNet	95.21	92.73	93.13	66.00	76.00	97.00	89.00	**92.00**	83.00	79.11	**77.51**	63.70	83.70
UNTER	96.80	92.13	**94.11**	75.00	77.00	**97.11**	91.33	89.05	84.71	78.81	76.77	74.11	85.58
Swin-UNet	90.66	79.61	83.28	66.53	55.77	94.29	76.60	85.47	75.52	70.33	67.58	61.93	75.63
HCI UNet	**99.54**	**93.92**	93.88	**93.50**	**77.18**	94.40	**95.42**	81.20	**88.40**	**84.40**	73.30	**80.40**	**87.96**

Table 3. Segmentation dice results of Synapse Dataset on multi-organ segmentation. The proposed method achieves state-of-the-art performance over the existing baseline. Note: Spl: spleen, RKid: right kidney, LKid: left kidney, Gall: gallbladder, Eso: esophagus, Liv: liver, Sto: stomach, Aor: aorta, IVC: inferior vena cava, Veins: portal and splenic veins, Pan: pancreas, AG: left and right adrenal glands.

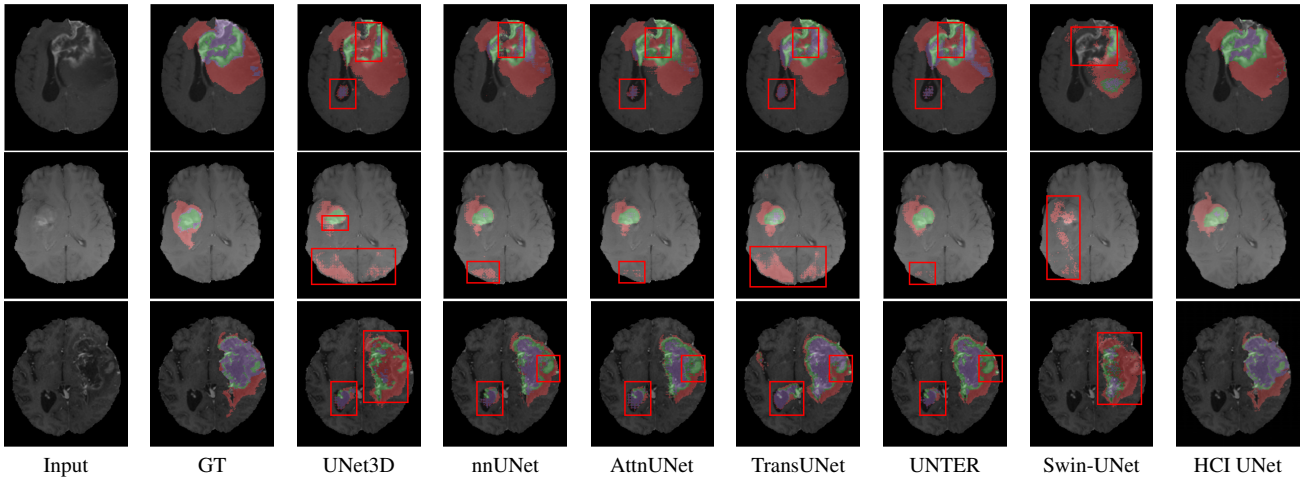


Figure 3. Qualitative comparison on the MSD Brain Dataset. GT denotes ground truth. Compared with all baseline methods, HCI UNet shows better segmentation performance, *i.e.* with less mis-segmentation, over-segmentation, or under-segmentation (as shown in the red box).

Model	#params	FLOPs (G)	BraTS DSC	BraTS HD95
UNet3D	1.981M	218.302	67.21	9.99
nnUNet	62.365M	8531.960	69.56	9.65
AttnUNet	5.909M	572.092	66.51	9.97
TransUNet	66.803M	2030.880	64.42	12.98
UNETR	92.618M	764.218	71.78	8.76
Swin-UNet	15.505M	774.392	64.36	11.89
HCI UNet	2.353M	311.808	78.07	7.96

Table 4. Computational cost on one 3D gray scale input with $(160, 160, 160)$. It is worth mentioning that TransUNet is a 2D framework. Thus, 3D input is split into 2D inputs and 2D outputs are merged into 3D output. For nnUNet, we follow the model configuration in [1] with deep supervision [16].

put channel as $(16, 32, 64, 128, 256)$ respectively. To note that TransUNet is a 2D solution, so we split the 3D input into 2D input, where the input size of the TransUNet is $(160, 160)$, and then merge all 2D output into 3D. Since nnUNet has a unique configuration, we keep the nnUNet configuration as the official application.

The experimental results of the computational cost are presented in Table 4 and Figure 5. An initial observation reveals a comparative enhancement in parameter efficiency across all techniques when juxtaposed with the baseline UNet3D model. Notably, Transformer-based approaches exhibit a substantial escalation in parameters, in contrast to the HCI UNet, which demonstrates a markedly lower parameter increment compared to Attention UNet. This disparity suggests that the marginal improvement in segmentation efficacy offered by Transformer methods may not justify their associated heightened computational demands.

The HCI UNet's position in Figure 5 indicates a favourable balance between the number of parameters, computational cost, and performance on BraTS DSC, which could suggest that it is a highly optimized model for medical image segmentation tasks in terms of both efficiency and accuracy. The model's performance, especially in comparison to other models with either many more parameters or significantly higher computational costs, may make it a compelling choice for practical applications where resources are limited but high accuracy is required.

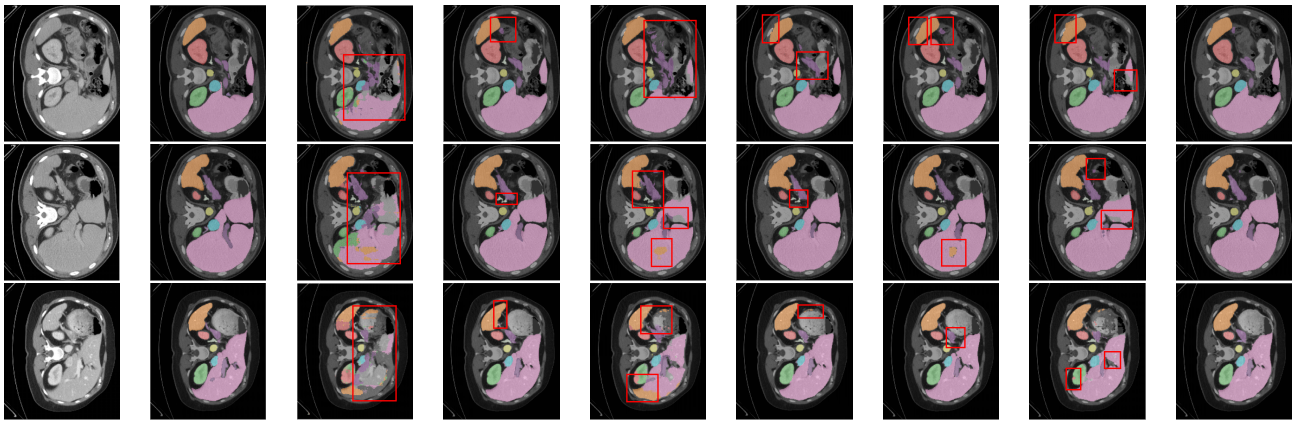
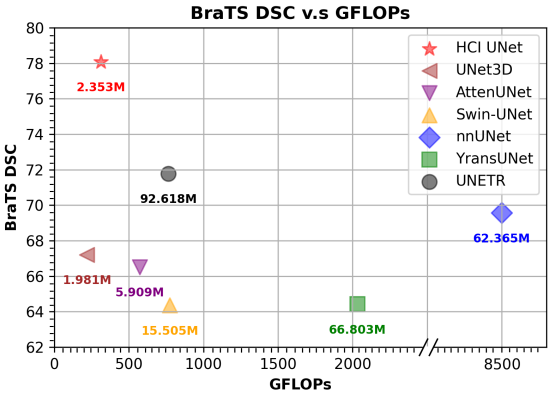
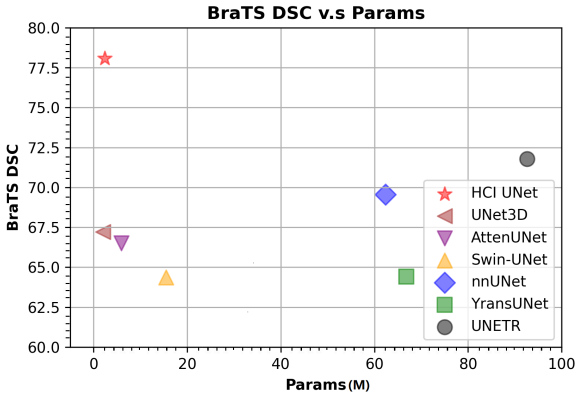


Figure 4. Qualitative comparison on the Synapse Dataset. GT denotes ground truth. Compared with all baseline methods, HCI UNet shows better segmentation performance, *i.e.* with less mis-segmentation, over-segmentation, or under-segmentation (as shown in the red box).



(a) Computational cost versus accuracy (DSC) on BraTS with $(160, 160, 160)$ input shape.



(b) Number of parameters versus accuracy (DSC) on BraTS with $(160, 160, 160)$ input shape.

Figure 5. Our simple HCI UNet (indicated by a red star) achieves the highest BraTS DSC score of 78.07 while having a moderate computational cost (311.808 GFLOPs) and parameters size ($2.353M$), while other methods including nnUNet, UNETR have either higher computational costs or lower performance metrics compared to HCI UNet.

6. Conclusion

In this paper, we propose a novel Hierarchical Context Interaction module designed to enhance U-Net architectures specifically for 3D medical image segmentation. Our findings indicate that the augmentation of U-Net structures with soft attention or self-attention mechanisms is not the sole pathway to performance improvement. Through extensive experiments on two 3D medical image segmentation benchmarks, namely the MSD Dataset and the Synapse Dateset, we demonstrate that our proposed HCI module can achieve competitive results with state-of-the-art methods in both terms of performance and efficiency.

Looking ahead, there are multiple promising directions for future research. Firstly, it would be beneficial to explore the adaptability of the HCI module to a broader range of neural network architectures beyond the U-Net framework. The HCI module should be recognized not just as an alternative to attention mechanisms but as a parameter-efficient tool. This is particularly significant in the context of medical image segmentation, where model efficiency and accuracy are paramount. Moreover, investigating the scalability of HCI modules, particularly in handling large-scale datasets and high-resolution images, is essential. The robustness and flexibility of HCI in these contexts are pivotal for its widespread deployment in diverse computer vision scenarios. This study provides insights into the generalization ability and efficacy of HCI modules across various models, potentially leading to broader applications in the field of image segmentation.

References

[1] Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al. The medical segmentation decathlon. *Nature Communications*, 13(1):4128, 2022. 2, 4, 5, 6, 7

[2] Reza Azad, Ehsan Khodapanah Aghdam, Amelie Rauland, Yiwei Jia, Atlas Haddadi Avval, Afshin Bozorgpour, Sanaz Karimijafarbigloo, Joseph Paul Cohen, Ehsan Adeli, and Dorit Merhof. Medical image segmentation review: The success of u-net. *arXiv preprint arXiv:2211.14830*, 2022. 1

[3] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *ArXiv Preprint, ArXiv:1607.06450*, 2016. 3, 4

[4] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-UNet: Unet-like pure transformer for medical image segmentation. In *ECCV Workshops*, pages 205–218. Springer, 2023. 2, 3, 4, 5

[5] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. TransUNet: Transformers make strong encoders for medical image segmentation. *ArXiv Preprint, ArXiv:2102.04306*, 2021. 5

[6] Sharan Chetlur, Cliff Woolley, Philippe Vandermersch, Jonathan Cohen, John Tran, Bryan Catanzaro, and Evan Shelhamer. cudnn: Efficient primitives for deep learning. *arXiv preprint arXiv:1410.0759*, 2014. 4

[7] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *CVPR*, pages 1251–1258, 2017. 3, 4

[8] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In *MICCAI*, pages 424–432. Springer, 2016. 5, 6

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009. 5

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv Preprint, ArXiv:2010.11929*, 2020. 3

[11] Tongle Fan, Guanglei Wang, Yan Li, and Hongrui Wang. Ma-net: A multi-scale attention network for liver and tumor segmentation. *IEEE Access*, 8:179656–179665, 2020. 2

[12] Jinjin Hai, Kai Qiao, Jian Chen, Hongna Tan, Jingbo Xu, Lei Zeng, Dapeng Shi, and Bin Yan. Fully convolutional densenet with multiscale context for automated breast tumor segmentation. *Journal of Healthcare Engineering*, 2019, 2019. 2

[13] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. UNETR: Transformers for 3d medical image segmentation. In *WACV*, pages 574–584, 2022. 5

[14] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37:1904–1916, 2014. 2

[15] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, pages 448–456. pmlr, 2015. 3, 4

[16] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021. 5, 7

[17] Sebastian Jaszczur, Aakanksha Chowdhery, Afroz Mohiuddin, Lukasz Kaiser, Wojciech Gajewski, Henryk Michalewski, and Jonni Kanerva. Sparse is enough in scaling transformers. *Advances in Neural Information Processing Systems*, 34:9895–9907, 2021. 4

[18] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ICLR*, 2015. 5

[19] Siyuan Li, Zedong Wang, Zicheng Liu, Cheng Tan, Haitao Lin, Di Wu, Zhiyuan Chen, Jiangbin Zheng, and Stan Z. Li. Efficient multi-order gated aggregation network. *ArXiv*, abs/2211.03295, 2022. 2

[20] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. *ArXiv Preprint, ArXiv:1312.4400*, 2013. 4

[21] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 10012–10022, 2021. 2

[22] Zhuang Liu, Hanzi Mao, Chaozheng Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11966–11976, 2022. 2

[23] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *3DV*, pages 565–571. IEEE, 2016. 5

[24] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention U-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018. 1, 2, 3, 5, 6

[25] Sahadev Poudel and Sang-Woong Lee. Deep multi-scale attentional features for medical image segmentation. *Applied Soft Computing*, 109:107445, 2021. 2

[26] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241, 2015. 1, 2

[27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017. 3

[28] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164, 2017. 1

[29] Lianghui Xu, Liejun Wang, Yongming Li, and Anyu Du. Big model and small model: Remote modeling and local information extraction module for medical image segmentation. *Applied Soft Computing*, 136:110128, 2023. 2

[30] Ping Xuan, Xixi Wu, Hui Cui, Qiangguo Jin, Linlin Wang, Tiangang Zhang, Toshiya Nakaguchi, and Henry B.L. Duh. Multi-scale random walk driven adaptive graph neural network with dual-head neighboring node attention for ct segmentation. *Applied Soft Computing*, 133:109905, 2023. 5

[31] Jianwei Yang, Chunyuan Li, and Jianfeng Gao. Focal modulation networks. *ArXiv*, abs/2203.11926, 2022. 2

[32] Fan Zhang, Yu Wang, and Hua Yang. Efficient context-aware network for abdominal multi-organ segmentation. *ArXiv Preprint, ArXiv:2109.10601*, 2021. 5

[33] Yichi Zhang, Qingcheng Liao, Le Ding, and Jicong Zhang. Bridging 2d and 3d segmentation networks for computation-efficient volumetric medical image segmentation: An empirical study of 2.5 d solutions. *Computerized Medical Imaging and Graphics*, 99:102088, 2022. 1