

# D-Former: A U-shaped Dilated Transformer for 3D Medical Image Segmentation

Yixuan Wu<sup>1</sup>, Kuanlun Liao<sup>2</sup>, Jintai Chen<sup>2</sup>, Jinhong Wang<sup>2</sup>, Danny Z. Chen<sup>3</sup>, Honghao Gao<sup>4,5\*</sup> and Jian Wu<sup>6\*</sup>

<sup>1</sup>School of Medicine, Zhejiang University, Hangzhou, 310030, China.

<sup>2</sup>College of Computer Science and Technology, Zhejiang University, Hangzhou, 310058, China.

<sup>3</sup>Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN 46556, USA.

<sup>4</sup>College of Future Industry, Gachon University, Seongnam, 13120, Korea.

<sup>5</sup>School of Computer Engineering and Science, Shanghai University, Shanghai, 200444, China.

<sup>6</sup>Second Affiliated Hospital School of Medicine and School of Public Health, Zhejiang University, Hangzhou, 310030, China.

\*Corresponding author(s). E-mail(s): [honghaogao@gachon.ac.kr](mailto:honghaogao@gachon.ac.kr); [gaohonghao@shu.edu.cn](mailto:gaohonghao@shu.edu.cn); [wujian2000@zju.edu.cn](mailto:wujian2000@zju.edu.cn);

Contributing authors: [wyx\\_chloe@zju.edu.cn](mailto:wyx_chloe@zju.edu.cn); [stevekl@zju.edu.cn](mailto:stevekl@zju.edu.cn); [jtchen721@gmail.com](mailto:jtchen721@gmail.com); [wangjinhong@zju.edu.cn](mailto:wangjinhong@zju.edu.cn); [dchen@nd.edu](mailto:dchen@nd.edu);

## Abstract

Computer-aided medical image segmentation has been applied widely in diagnosis and treatment to obtain clinically useful information of shapes and volumes of target organs and tissues. In the past several years, convolutional neural network (CNN) based methods (e.g., U-Net) have dominated this area, but still suffered from inadequate long-range information capturing. Hence, recent work presented computer vision Transformer variants for medical image segmentation tasks and obtained promising performances. Such Transformers model long-range dependency by computing pair-wise patch relations. However, they incur prohibitive computational costs, especially on 3D medical

images (e.g., CT and MRI). In this paper, we propose a new method called Dilated Transformer, which conducts self-attention for pair-wise patch relations captured alternately in local and global scopes. Inspired by dilated convolution kernels, we conduct the global self-attention in a dilated manner, enlarging receptive fields without increasing the patches involved and thus reducing computational costs. Based on this design of Dilated Transformer, we construct a U-shaped encoder-decoder hierarchical architecture called D-Former for 3D medical image segmentation. Experiments on the Synapse and ACDC datasets show that our D-Former model, trained from scratch, outperforms various competitive CNN-based or Transformer-based segmentation models at a low computational cost without time-consuming per-training process.

**Keywords:** Medical Image Analysis, Segmentation, Transformer, Long-range Dependency, Position Encoding.

## 1 Introduction

Medical image segmentation, as one of the critical computer-aided medical image analysis problems, aims to capture precisely the shapes and volumes of target organs and tissues by pixel-wise classification, obtaining clinically useful information for diagnosis, treatment, and intervention. With the recent development of deep learning methods and computer vision algorithms, medical image segmentation has been revolutionized and achieved remarkable progresses (e.g., automatic liver and tumor lesion segmentation [66], brain tumor segmentation [68], and multiple sclerosis (MS) lesion segmentation [67]).

Fully convolutional network (FCN) [18] was first proved effective for general image segmentation tasks, which became a predominant technique for medical image segmentation [55–59]. However, it was observed that vital details can be missing with the decrease of feature map sizes when FCN models go deeper. To this end, a family of U-shaped networks [20, 60–65, 77] extended the sequential FCN frameworks into encoder-decoder type architectures, alleviating the spatial information loss using skip-connections. Competitive solutions were given by DeepLab models [73–76], which applied atrous convolutions instead of pool layers to expand the receptive field and introduced fully connected conditional random field (CRF) to maintain fine details. Although these CNN-based methods have achieved great performances in medical image segmentation tasks, they still suffered from limited receptive fields and cannot capture long-range dependency, leading to sub-optimal accuracy and failing to meet the needs of various medical image segmentation scenarios.

Inspired by the success of Transformer with its self-attention mechanism in natural language processing (NLP) tasks [34, 36], researchers tried to adapt Transformers [35, 37, 39, 40] to computer vision in order to compensate the locality of CNNs. The self-attention mechanism in Transformers is able to compute pair-wise relations between patches globally, consequently achieving

feature interactions across a long range. Non-local neural network [41] was the first to adopt self-attention mechanism to complement CNNs for modeling pixel-level long-range dependency for visual recognition tasks. Then, the Vision Transformer (ViT) [35] utilized a pure Transformer framework to deal with vision tasks, treating an image as a collection of spatial patches. Recently, Transformers have achieved excellent outcomes on a variety of vision tasks [13–15, 42, 45, 46], including image recognition [15, 45–48], semantic segmentation [42], and object detection [13, 14]. On semantic medical image segmentation, Transformer-combined architectures can be divided into two categories: the main one adopts self-attention like operations to complement CNNs [1, 49–51]; the other uses pure Transformers to constitute encoder-decoder architectures so as to capture deep representations and predict the class of each image pixel [42–44, 53].

Although the above medical image segmentation methods were promising and yielded good performance to some extent, they still suffered considerable drawbacks. (1) The majority of these Transformer segmentation models were designed for 2D images [1, 43, 44, 51, 53]. For 3D medical images (e.g., 3D MRI scans), they may divide the input images into 2D slices and process the individual slices with the 2D models, which could lose useful 3D contextual information. (2) Compared with common 2D natural scene images, processing 3D medical images inevitably incurs larger model sizes and computation costs, especially when computing global feature interactions with self-attention in vanilla Transformer [34] (see more details in Sec. 3.3). Although some adaptations were proposed to reduce the operation scopes of self-attention [15–17, 45, 46, 52, 54] (e.g., Pyramid Vision Transformer [45] used progressive scaling pyramids to reduce the computation of large feature maps), they could lead to insufficient global information fusion. (3) The self-attention operation in Transformers was shown to be permutation-equivalence [34], which omits the order of patches in an input sequence. However, the permutation-equivalence nature can be detrimental to medical image segmentation since segmentation results are often highly position-correlated. Prior work usually utilized absolute position encoding (APE) [34] or relative position encoding (RPE) [46, 69] to supplement position information. But, APE requires a pre-given and fixed patch amount and thus fails to generalize to different image sizes, while RPE ignores the absolute position information that can be a vital cue in medical images (e.g., the positions of bones are often relatively stable).

To address the drawbacks above, we propose a new efficient model called **Dilated Transformer** (D-Former) to directly process 3D medical images (instead of dealing with 2D slices of 3D images independently) and predict volumetric segmentation masks. Our proposed D-Former is a 3D U-shaped architecture with hierarchical layers, and employs skip connections from encoder to decoder following [20, 60–65, 77]. This model’s stem is constructed with eight D-Former blocks, each of which consists of several local scope modules (LSMs) and global scope modules (GSMs). The LSM conducts self-attention locally, focusing on fine information capturing. The GSM performs

global self-attention on uniformly sampled patches, aiming to explore rough and long-range dependent information at low cost. The LSMs and GSMs are arranged in an alternate manner to achieve local and global information interaction. For drawback (3) above, we manage to incorporate position information among patches in a more dynamic manner. Inspired by [24, 52], we utilize depth-wise convolutions [80] to learn position information, which can help provide useful position cues in medical image segmentation.

Benefiting from these designs, our proposed D-Former model could be more suitable for medical image segmentation tasks and yield better segmentation accuracy. The main contributions of this work are as follows:

- (1) We construct a 3D Transformer based architecture which allows to process volumetric medical images as a whole and thus spatial information along the depth dimension of 3D medical images to be fully captured.
- (2) We design local scope modules (LSMs) and global scope modules (GSMs) to increase the scopes of information interactions without increasing the patches involved in computing self-attention, which helps reduce computational costs.
- (3) To further incorporate relative and absolute position information among patches, we apply a dynamic position encoding method to learn it from the input directly. As a result, an inherent problem of common Transformers, permutation-equivalence [34], could be considerably alleviated.
- (4) Extensive experimental evaluations show that our model outperforms state-of-the-art segmentation methods in different domains (e.g., CT and MRI), with smaller model sizes and less FLOPs than the known methods.

## 2 Related Work

### 2.1 CNN-based Segmentation Networks

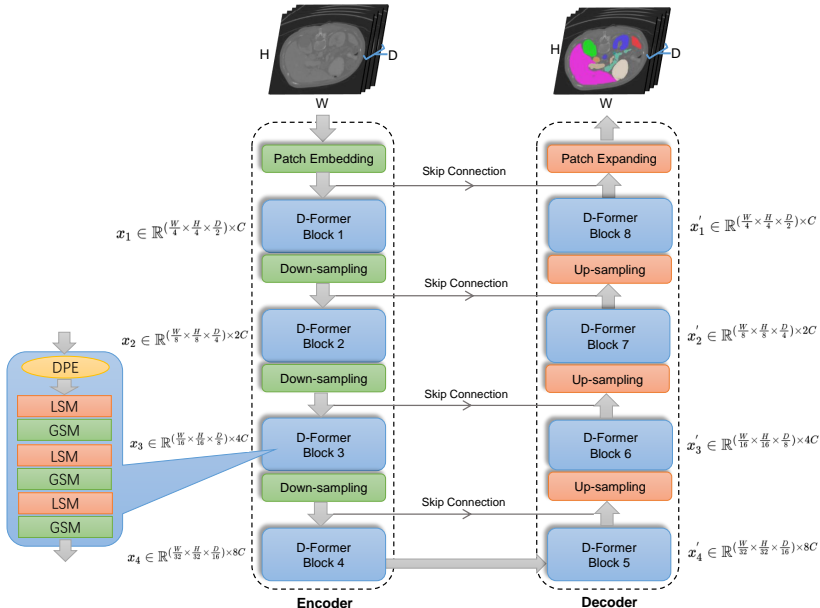
Since the advent of the seminal U-Net model [20], many CNN-based networks have been developed [30, 32, 33, 70, 77]. As for the design of skip connections, U-Net++ [29] and U-Net3+ [19] were proposed to attain dense connections between encoder and decoder. In addition, regarding the locality of CNNs, researchers have designed different kinds of mechanisms to enlarge the receptive field, such as larger kernel [2], dilated convolution module [4, 5], pyramid pooling module [8, 9], and deformable convolution module [11, 12]. In particular, dilated convolution is an ingenious design, which expands the convolution kernel by inserting holes between its consecutive elements. This design has been adopted by various segmentation models, achieving good performance compared with the original convolution-based methods. Our Dilated Transformer also obtains a key idea from this design and aims to conduct self-attention in a patch skipping manner (see Sec. 3.3 for details).

## 2.2 Visual Transformer Variants

Transformer and its self-attention mechanism were first designed for sequence modeling and transduction tasks in the domain of natural language processing (NLP), achieving state-of-the-art performance [34, 36]. Inspired by tremendous success of Transformer in NLP, researchers adapted Transformer to computer vision tasks. The first attempt was vision Transformer (ViT) [35], which needed huge pre-training datasets. In order to overcome this weakness, DeiT [40] was proposed, and this work provided a wide range of training strategies with knowledge distillation, which contributed to better performance of vanilla Transformer. There were different kinds of adaptations for vanilla Transformer, such as Swin Transformer [46], pyramid vision Transformer [45], Transformer in Transformer [71], and aggregating nested Transformers [15]. In particular, Swin Transformer showed great success in various computer vision tasks with its elegant shift window mechanism and hierarchical architecture. Our proposed D-Former is inspired by Swin Transformer’s local-global combining scopes of information interactions.

## 2.3 Transformers for Segmentation Tasks

As mentioned above, Transformers used in medical image segmentation methods can be divided into two categories. In the main category, Transformer and its self-attention mechanism are utilized as a supplement for the convolution-based stem. SETR [38] was proposed to apply Transformer as encoder to extract features for segmentation tasks. In medical imaging, many models with Transformers focused on segmentation tasks. In TransUNet [1], convolutional layer was used as a feature extractor to obtain detailed information from raw images; it then generated feature maps which are put into Transformer layer to obtain global information. UNETR [49] proposed a 3D Transformer-combining architecture for medical images, which treated Transformer layer as encoder to extract features and convolutional layer as decoder. A great amount of such work focused on taking advantage of both Transformer’s long-range dependency and CNN’s inductive bias. In the other category, Transformer is regarded as the main stem for building the whole architecture [42–44, 53]. MedT [72] proposed the Gated Axial-Attention mechanism and used Gated Axial Transformer layer to build the whole architecture. Swin-Unet [53] used the Swin Transformer block as a basic unit to build a U-shape architecture with skip connections. DS-TransUNet [44] further extended Swin-Unet by adding another encoder pathway for input of different sizes. nnFormer [42] modified the up-sampling and down-sampling blocks with convolutions based on the Swin-Unet architecture.



**Fig. 1:** The overall architecture of our D-Former model. Each D-Former block is constructed with one dynamic position encoding block (DPE) and several local scope modules (LSMs) and global scope modules (GSMs). The input size of the D-Former block  $i$  is reported sideward, and the output sizes are the same as the corresponding input sizes. The values in round brackets denote the numbers of patches, which are regarded as one dimension when computed in Transformers (i.e.,  $(\frac{W}{4} \times \frac{H}{4} \times \frac{D}{2})$ ,  $(\frac{W}{8} \times \frac{H}{8} \times \frac{D}{4})$ ,  $(\frac{W}{16} \times \frac{H}{16} \times \frac{D}{8})$ ,  $(\frac{W}{32} \times \frac{H}{32} \times \frac{D}{16})$ ).

## 3 Method

### 3.1 The Overall Architecture

Our proposed D-Former model is outlined in Fig. 1, which is a hierarchical encoder-decoder architecture. The encoder pathway consists of one patch embedding layer for transforming 3D images into sequences and four proposed D-Former blocks for feature extraction with three down-sampling layers in between them. The first, second, and fourth D-Former blocks each consist of one local scope module (LSM) and one global scope module (GSM) respectively, while the third D-Former block has three LSMs and three GSMs, in which the LSMs and GSMs are arranged in an alternate manner. The decoder pathway is symmetric to the encoder pathway, which also has four D-Former blocks, three up-sampling layers, and one patch expanding layer. In addition, skip connections are used to transfer information from the encoder to the decoder at the corresponding levels. The feature maps from the encoder are

concatenated with the corresponding feature maps along the channel dimension, which may compensate for the loss of fine-grained information as the model goes deep.

In this section below, we will present the components of **D-Former** one by one, including the patch embedding and patch expending layers (Sec. 3.2), the D-Former block and its major modules, the local scope module and global scope module (Sec. 3.3), the down-sampling and up-sampling operations (Sec. 3.4), and the dynamic position encoding block (Sec. 3.5).

## 3.2 Patch Embedding & Patch Expending

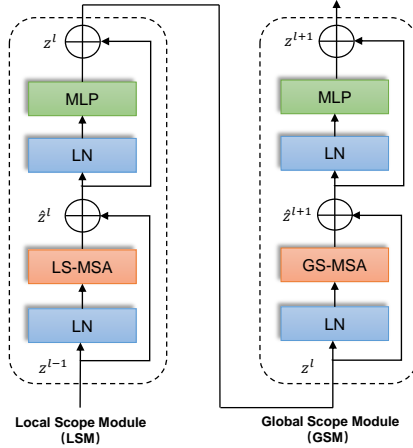
Similar to common Transformers in computer vision, after data augmentation, an input 3D medical image  $x \in \mathbb{R}^{W \times H \times D}$  first goes through a patch embedding layer and is divided into a series of patches of size  $4 \times 4 \times 2$  each, and then is projected into  $C$  channel dimensions by linear projection to yield a feature map (denoted by  $x_1$ ) of size  $(\frac{W}{4} \times \frac{H}{4} \times \frac{D}{2}) \times C$ , where  $(\frac{W}{4} \times \frac{H}{4} \times \frac{D}{2})$  denotes the number of patches and  $C$  is the number of the channel dimensions. Hence, the input 3D image is reorganized as a sequence (of length  $(\frac{W}{4} \times \frac{H}{4} \times \frac{D}{2})$ ) and can be directly fed to a Transformer architecture. The final patch expending layer is used to restore the feature map to the original input size, and utilizes a segmentation head (like 3D UNet [31]) to yield pixel-wise segmentation masks.

## 3.3 D-Former Blocks

After patch embedding,  $x_1$  is directly fed to D-Former block 1. In the processing by Transformer block 1,  $x_1$  is first processed by a new dynamic position encoding block that embeds position information into feature maps (see details in Sec. 3.5), and then it is operated by the Local Scope Module (LSM) and Global Scope Module (GSM) alternatively to extract higher-level features. The other D-Former Blocks process the corresponding input features similarly, and the feature map sizes are provided in Fig. 1.

### 3.3.1 Local Scope Module and Global Scope Module

The local scope module (LSM) and global scope module (GSM) are presented to capture local and global features respectively, which employ two different self-attention operations, called local scope multi-head self-attention (LS-MSA) and global scope multi-head self-attention (GS-MSA). As shown in Fig. 2, an LSM is composed of a LayerNorm layer [79], a proposed LS-MSA, another LayerNorm layer, and a Multilayer Perceptron (MLP), in sequence, with two residual connections to prevent gradient vanishing [34]. In a GSM, the LS-MSA is replaced by a proposed GS-MSA, and the other components are kept the same as the LSM. To allow local features and global features to be captured and fused well, we arrange the LSM and GSM alternatively in each D-Former block. With these components, their operations are formally



**Fig. 2:** The local scope module (LSM) and global scope module (GSM), which should be arranged in pair to combine local and global information.

defined as:

$$\hat{z}^l = LS\text{-MSA}(LN(z^{l-1})) + z^{l-1}, \quad (1)$$

$$z^l = MLP(LN(\hat{z}^l)) + \hat{z}^l, \quad (2)$$

$$\hat{z}^{l+1} = GS\text{-MSA}(LN(z^l)) + z^l, \quad (3)$$

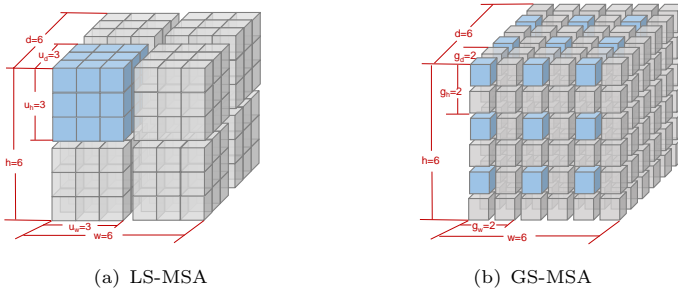
$$z^{l+1} = MLP(LN(\hat{z}^{l+1})) + \hat{z}^{l+1}, \quad (4)$$

where  $\hat{z}^l$  and  $z^l$  denote the outputs of LS-MSA and the corresponding MLP respectively, and  $\hat{z}^{l+1}$  and  $z^{l+1}$  denote the outputs of GS-MSA and the corresponding MLP respectively.

### 3.3.2 Local Scope Multi-head Self-Attention (LS-MSA)

A vanilla Transformer conducts self-attention in a global scope in order to capture pair-wise relationships between patches, leading to quadratic complexity with respect to the number of patches. However, due to the fact that 3D medical images would increase computation inevitably, this original self-attention would not be suitable for 3D medical image related tasks, especially for semantic segmentation with dense prediction targets. Under such circumstances, as illustrated in Fig. 3(a), we first evenly divide a whole feature map into non-overlapping units (we denote the number of patches in each unit by  $u_d \times u_h \times u_w$ , where  $u_d$  denotes the number of patches in one unit along the depth dimension  $D$ ,  $u_h$  along the height dimension  $H$ , and  $u_w$  along the width dimension  $W$ ), and conduct self-attention within each unit. In this way, the computational complexity will be reduced to linear in terms of the number of patches in the whole feature map. The computational complexity ( $\Omega$ ) of these





**Fig. 3:** (a) The local scope multi-head self-attention: The self-attention is conducted in a local unit (colored in blue) where the patches are adjacent. (b) The global scope multi-head self-attention: The self-attention is conducted in a global unit (colored in blue) where patches are picked every  $g$ -th patch across the feature map. A small cube represents one patch. The feature map size is set as  $6 \times 6 \times 6$  and the unit size is  $3 \times 3 \times 3$  as an example. We color only the patches of one unit in blue for illustration; the other grey patches are also utilized to construct seven other units in both LS-MSA and GS-MSA.

two different self-attention mechanisms are computed as:

$$\Omega(MSA) = 4dhwC^2 + 2(dhw)^2C, \quad (5)$$

$$\Omega(LS-MSA) = 4dhwC^2 + 2u_du_hu_wdhwC, \quad (6)$$

where  $u_du_hu_w$  denotes the number of patches in one unit, and  $dhw$  denotes the number of patches in the whole feature map ( $d$ ,  $h$ , and  $w$  denote the depth, height, and width of the feature map, respectively). In most cases,  $u_du_hu_w \ll dhw$ . The *Softmax* operation is omitted when computing the computational complexity.

### 3.3.3 Global Scope Multi-head Self-Attention (GS-MSA)

The LS-MSA performs self-attention only within each local unit, which lacks global information interaction and long-rang dependency. To address this issue, we design a global scope multi-head self-attention mechanism to attain information interaction across different units in a dilated manner. As illustrated in Fig. 3(b), for a whole feature map, we pick one patch every  $g$  distance along each dimension and form a unit with all the patches thus picked, on which self-attention would then be conducted. Likewise, we pick the other patches to form new units, until all the patches are utilized. Hence, the receptive field in computing self-attention will be enlarged but the number of patches involved will not be increased, which means that it would not increase the computational cost while getting access to long-range information interaction. To keep consistency between LSM and GSM, we set  $d = g_d \times u_d$ ,  $h = g_h \times u_h$ , and

$w = g_w \times u_w$ , which ensure that the numbers of units in LSM and GSM are kept the same. Here,  $d \times h \times w$  denotes the number of patches in the whole feature map,  $u_d \times u_h \times u_w$  denotes the number of patches in one unit, and  $g_d$ ,  $g_h$ , and  $g_w$  denote the distance between two nearest patches picked along the depth dimension  $D$ , height dimension  $H$ , and width dimension  $W$ , respectively.

### 3.4 Down-sampling & Up-sampling

Between every two adjacent D-Former blocks of the encoder, a down-sampling layer is utilized to merge patches for further feature fusion. Specifically, a down-sampling layer concatenates the feature maps of  $2 \times 2 \times 2$  neighboring patches (2 neighboring patches along the width, height, and depth dimensions, respectively), reducing the number of patches by 8 times. Then, a fully connected layer is utilized to reduce the feature channel size by 4 times to ensure that the channel size can be doubled after each down-sampling layer. Thus, the output feature maps of each down-sampling layer will be  $x_2 \in \mathbb{R}^{(\frac{W}{8} \times \frac{H}{8} \times \frac{D}{4}) \times 2C}$ ,  $x_3 \in \mathbb{R}^{(\frac{W}{16} \times \frac{H}{16} \times \frac{D}{8}) \times 4C}$ , and  $x_4 \in \mathbb{R}^{(\frac{W}{32} \times \frac{H}{32} \times \frac{D}{16}) \times 8C}$ , respectively. In reverse to the down-sampling layers, four up-sampling layers of the decoder are used to enlarge the low-resolution feature maps and reduce the number  $C$  of the channel dimensions. In this way, our model will be able to extract features in a multi-scale manner and yield better segmentation accuracy.

### 3.5 The Dynamic Position Encoding Block

The depth-wise convolution (*DW-Conv*) is a type of convolution that applies a single convolutional filter for each input channel instead of for all channels as in a common convolution, which can decrease the computational cost. We apply 3D depth-wise convolution [80] to the input feature maps (or images) once in every D-Former block to learn position information. Then the learned position information will be added to the original input  $x_i$  as:

$$x'_i = \text{Resize}(\text{DW-Conv}(\text{Resize}(x_i))) + x_i, \quad (7)$$

where  $x_i$  denotes the input feature maps of the  $i$ -th D-Former block and  $x'_i$  denotes the output feature maps embedded with position information. *Resize* is used to adjust the dimensions of feature maps  $x_i$  to cater the input need of DW-Convolution.

In this way, position information among patches can be extracted by a DW-Convolution. Given the fact that position information could be dynamically learned based on the input  $x$  itself, a drawback in the previous work that requires a fixed number of patches can be avoided. In addition, the convolution’s inherent nature of translation-invariance can be utilized to increase the stability and generalization performance [25].

## 4 Experiments

### 4.1 Datasets

The **Synapse multi-organ segmentation (Synapse)** dataset includes 30 axial contrast-enhanced abdominal CT scans. Following the training-test split in [1], 18 of the 30 scans are used for training and the remaining ones are for testing. We take the average Dice Similarity Coefficient (DSC) [77] as the measure for evaluating the segmentation performances of eight target organs, including aorta, gallbladder, kidney (L), kidney (R), liver, pancreas, spleen, and stomach.

The **Automated Cardiac Diagnosis Challenge (ACDC)** dataset contains 150 magnetic resonance imaging (MRI) 3D cases collected from different patients, and each case covers a heart organ from the base to the apex of the left ventricle. Following the setting in [1], only 100 well-annotated cases are used in the experiments, and the training, validation, and test data are partitioned with the ratio of 7: 1: 2. For fair comparison, the average DSC is employed to evaluate the segmentation performances following the previous work [1], and three key parts of the heart are chosen, including the right ventricle (RV), myocardium (Myo), and left ventricle (LV).

### 4.2 Implementation Setup

**Pre-training.** Our D-Former model is trained from scratch, which means that we initialize the model’s weights randomly. Note that in common practice, pre-training is important to Transformer-based models. This is because the pre-training process provides generalized representations and prior knowledge for down-stream tasks. For example, in vision Transformer (ViT) [35], it is considered that the model performance depends heavily on pre-training, and its experiments verified this view. Besides, lots of known medical image segmentation methods used pre-trained weights to initialize their models [1, 27, 28, 42, 51]. However, the pre-training process of Transformer-based models brings up two issues. First, the pre-training process usually incurs high computational complexity in terms of time or computation consumed. Second, for medical images, there are few complete and acknowledged sizable datasets for pre-training (in comparison, ImageNet [81] is available for natural scene images), and the domain gap between natural images and medical images makes it hard for medical image segmentation models to use existing large natural image datasets directly. For these reasons, we choose to train our D-Former model from scratch, which nevertheless yields promising performance that surpasses state-of-the-art methods with pre-training.

**Implementation details.** Our proposed D-Former is implemented on PyTorch 1.8.0, and all the experiments are trained on an NVIDIA GeForce RTX 3090 GPU with 24 GB memory. The batch size during training is 2 and during inference is 1. The SGD optimizer [82] with momentum 0.99 is used.

**Table 1:** Segmentation performances of different methods on the Synapse dataset (average Dice Similarity Coefficient (DSC) in %).

Method	Average	Aotra	Gallbladder	Kidnery (L)	Kidnery (R)	Liver	Pancreas	Spleen	Stomach
V-Net	68.81	75.34	51.87	77.10	80.75	87.84	40.04	80.56	56.98
DARR	69.77	74.74	53.77	72.31	73.24	94.08	54.18	89.90	45.96
R50 U-Net	74.68	87.74	63.66	80.60	78.19	93.74	56.90	85.87	74.16
R50 Att-UNet	75.57	55.92	63.91	79.20	72.71	93.56	49.37	87.19	74.95
U-Net	76.85	89.07	69.72	77.77	68.60	93.43	53.98	86.67	75.58
Att-UNet	77.77	89.55	68.88	77.98	71.11	93.57	58.04	87.30	75.75
VIT	67.86	70.19	45.10	74.70	67.40	91.32	42.00	81.75	70.44
R50 VIT	71.29	73.73	55.13	75.80	72.20	91.51	45.99	81.99	73.95
TransUNet	77.48	87.23	63.13	81.87	77.02	94.08	55.86	85.08	75.62
Swin-UNet	79.13	85.47	66.53	83.28	79.61	94.29	56.58	90.66	76.60
TransClaw U-Net	78.09	85.87	61.38	84.83	79.36	94.28	57.65	87.74	73.55
LeVit-Unet-384	78.53	87.33	62.23	84.61	80.25	93.11	59.07	88.86	72.76
nnFormer	87.40	92.04	71.09	87.64	87.34	96.53	<b>82.49</b>	92.91	<b>89.17</b>
MISSFormer	81.96	86.99	68.65	85.21	82.00	94.41	65.67	91.92	80.81
D-Former	<b>88.83</b>	<b>92.12</b>	<b>80.09</b>	<b>92.60</b>	<b>91.91</b>	<b>96.99</b>	76.67	<b>93.78</b>	86.44

\* The DSC values of the compared models are from open source of the original papers.

The initial learning rate is 0.01 with weight decay of  $3e-5$ . The poly learning rate strategy [21] is utilized with the maximum training epochs of 3000 for the Synapse dataset and 1500 for the ACDC dataset.

**Loss function.** The cross entropy loss and Dice loss are both widely used for general segmentation tasks. However, since the cross entropy loss is apt to perform well for uniform class distribution while Dice loss is more suitable for target objects of large sizes [23], each of them alone may not be effective for medical image segmentation tasks that involve imbalanced classes and target objects of small sizes. Thus, our loss function combines the binary cross entropy loss  $Y$  [22] and Dice loss  $\hat{Y}$  [77] together, which is defined as:

$$\mathcal{L}(Y, \hat{Y}) = -\frac{1}{N} \sum_{n=1}^N \left( \frac{1}{2} \cdot Y_n \cdot \log \hat{Y}_n + \frac{2 \cdot Y_n \cdot \hat{Y}_n}{Y_n + \hat{Y}_n} \right) \quad (8)$$

where  $Y_n$  and  $\hat{Y}_n$  denote the ground truth and predicted probabilities of the  $n^{th}$  image respectively, and  $N$  is the batch size.

### 4.3 Quantitative Results

We evaluate the performance of our proposed D-Former model on the Synapse and ACDC datasets, and compare with various state-of-the-art models, including V-Net [77], DARR [26], R50 U-Net [20], R50 Att-UNet [78], U-Net [20], Att-UNet [78], VIT [35], R50 VIT [35], TransUNet [1], Swin-UNet [53], LeVit-Unet-384 [28], nnFormer [42], and MISSFormer [43].

Quantitative results on the Synapse dataset are reported in Table 1, which show that our method outperforms the previous work by a clear margin. It is notable that the concurrent Transformer-based methods nnFormer and MISSFormer achieve some performance gains compared to the CNN-based

**Table 2:** Segmentation performances of different methods on the ACDC dataset (average Dice Similarity Coefficient (DSC) in %).

Method	Average	RV	Myo	LV
R50 U-Net	87.55	87.10	80.63	94.92
R50 Att-UNet	86.75	87.58	79.20	93.47
VIT	81.45	81.46	70.71	92.18
R50 VIT	87.57	86.07	81.88	94.75
TransUNet	89.71	88.86	84.54	95.73
Swin-UNet	90.00	88.55	85.62	95.83
LeVit-Unet-384	90.32	89.55	87.64	93.76
nnFormer	91.78	90.22	89.53	95.59
MISSFormer	87.90	86.36	85.75	91.59
D-Former	<b>92.29</b>	<b>91.33</b>	<b>89.60</b>	<b>95.93</b>

\* The DSC values of the compared models are from open source of the original papers.

methods, while our method still brings further improvement in the average DSC by 1.43% compared to nnFormer and by 6.87% compared to MISSFormer. Besides, our D-Former obtains accuracy improvement on almost every organ class, except for the pancreas and stomach, which verifies that our D-Former is a promising and robust framework.

Quantitative results on the ACDC dataset are reported in Table 2, and a similar conclusion can be drawn. D-Former achieves the best average DSC of 92.29% without pre-training. Compared with the other methods, our method brings improvements in the average DSC by 4.74% over R50 U-Net and by 5.54% over R50 Att-UNet. Compared to the concurrent Transformer-based methods, our method still achieves 0.51% performance gain over nnFormer and 4.39% over MISSFormer in the average DSC. Specifically, among all the key parts of the heart, including the right ventricle (RV), myocardium (Myo), and left ventricle (LV), our D-Former obtains the best segmentation accuracy compared to the other methods in the average DSC.

From the results in Table 1 and Table 2, one can see that our D-Former attains excellent generalization on both CT data and MRI data, outperforming the previous methods. Notably, different from most of the known Transformer based frameworks that require a pre-training process, D-Former is initialized randomly and is trained from scratch, yet still obtains competitive performances. This implies that our model could be more suitable for medical imaging tasks when general large size medical image pre-training datasets (such as ImageNet [81] for natural scene images) are lacking.

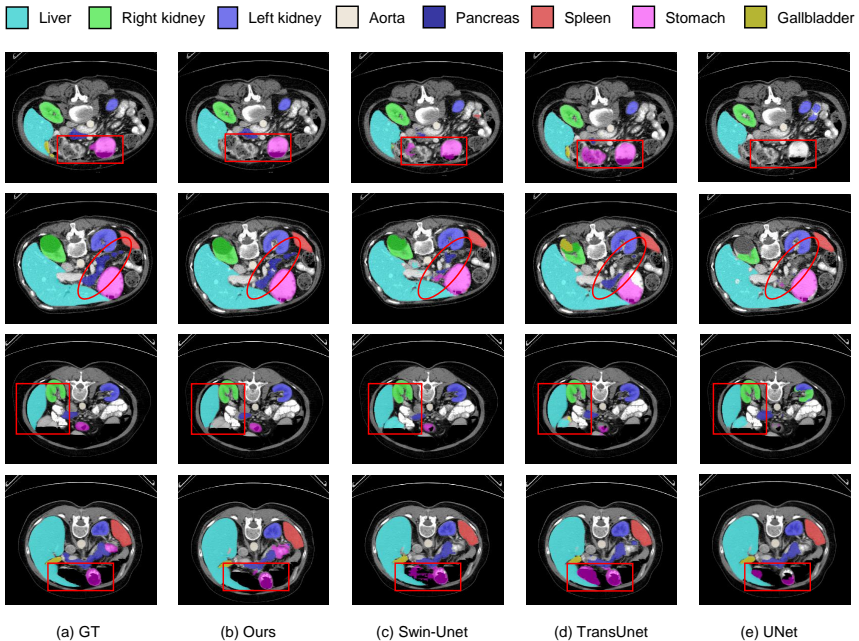
#### 4.4 Comparison of Model Complexity

In Table 3, we compare the numbers of parameters and floating point operations (FLOPs) of our proposed D-Former with those of different 3D medical image segmentation models, including UNETR [49], CoTr [50], TransBTS [27], and nnFormer [42]. The number of FLOPs is calculated based on the input image size of  $64 \times 128 \times 128$  for fair comparison. We should note that we omit the part of the complexity brought by activation functions and normalization

**Table 3:** Comparison of the numbers of parameters and FLOPs among various methods that segment 3D medical images directly.

Method	#Params (M)	FLOPs (G)
3D U-Net	16.31	947.69
UNETR	92.25	86.02
CoTr	41.86	377.48
TransBTS	32.19	171.30
nnFormer	158.92	157.88
D-Former	44.26	54.46

\* The numbers of FLOPs are computed with the input image size of  $D \times W \times H = 64 \times 128 \times 128$ .

**Fig. 4:** Visual comparison with several state-of-the-art methods on the Synapse dataset. The red-marked regions indicate that our model attains discriminative segmentation performance.

layers. From Table 3, one can see that our D-Former has 44.26M parameters and 54.46G FLOPs, which has a lower computational cost compared to nnFormer (157.88G FLOPs), TransBTS (171.30G FLOPs), CoTr (377.48G FLOPs), UNETR (86.02G FLOPs), and 3D U-Net (947.69G FLOPs). Moreover, even if CNN-based models (e.g., 3D U-Net) have less parameters, they are still burdened with high model complexity.

**Table 4:** Ablation study on the effect of the Global Scope Module (GSM) (average Dice Similarity Coefficient (DSC) in %).

Method	Average	Aotra	Gallbladder	Kidney (L)	Kidney (R)	Liver	Pancreas	Spleen	Stomach
w/o GSM	88.17	91.00	79.53	92.40	91.62	96.39	<b>77.68</b>	92.74	84.00
GSM (ours)	<b>88.83</b>	<b>92.12</b>	<b>80.09</b>	<b>92.60</b>	<b>91.91</b>	<b>96.99</b>	76.67	<b>93.78</b>	<b>86.44</b>

**Table 5:** Ablation study on the effect of the Global Scope Multi-head Self-attention (GS-MSA) (average Dice Similarity Coefficient (DSC) in %).

Method	Average	Aotra	Gallbladder	Kidney (L)	Kidney (R)	Liver	Pancreas	Spleen	Stomach
SW-MSA	87.50	91.37	77.48	92.55	<b>93.07</b>	96.26	76.62	88.36	84.27
GS-MSA (ours)	<b>88.83</b>	<b>92.12</b>	<b>80.09</b>	<b>92.60</b>	91.91	<b>96.99</b>	<b>76.67</b>	<b>93.78</b>	<b>86.44</b>

**Table 6:** Ablation study on the effect of dynamic position encoding (DPE) (average Dice Similarity Coefficient (DSC) in %).

Method	Average	Aotra	Gallbladder	Kidney (L)	Kidney (R)	Liver	Pancreas	Spleen	Stomach
APE	84.78	88.14	76.65	90.16	86.38	95.65	67.47	91.35	82.47
SPE	86.04	88.51	77.13	<b>93.75</b>	88.47	96.40	68.47	92.77	82.81
RPE	86.35	90.37	78.41	92.30	87.37	96.06	70.86	91.70	83.76
DPE (ours)	<b>88.83</b>	<b>92.12</b>	<b>80.09</b>	92.60	<b>91.91</b>	<b>96.99</b>	<b>76.67</b>	<b>93.78</b>	<b>86.44</b>

## 4.5 Qualitative Visualizations

To intuitively demonstrate the performances of our D-Former model, we compare some qualitative results of our model with several other methods (including Swin-Unet, TransUnet, and UNet) on the Synapse dataset in Fig. 4. One can see that the predicted organ masks of our model are much more similar to the ground truth in general. As for specific organs, our model has better accuracy in identifying and sketching the contours of stomach (e.g., the first and fourth rows), which is consistent with the conclusions based on the quantitative results above. In the second row, only our model can delineate the outline of pancreas well, thus suggesting that our model has a better ability to capture long-range dependency given the fact that the shape of pancreas is long and narrow. In addition, as illustrated in the third row, our D-Former is able to identify the true region of liver, while the other three models incur some mistakes on the liver. This shows that our method is effective at exploiting the relations between the target organs’ patches and the other patches, owing to our model’s dynamic position encoding block. In a nutshell, the qualitative visualizations convey an intuitive demonstration of our model’s high segmentation accuracy, especially on some slices that are difficult to segment.

## 4.6 Ablation Studies

We conduct ablation studies on the Synapse dataset to evaluate the effectiveness of our model design.

**Table 7:** Ablation study on the position of the dynamic position encoding block (average Dice Similarity Coefficient (DSC) in %).

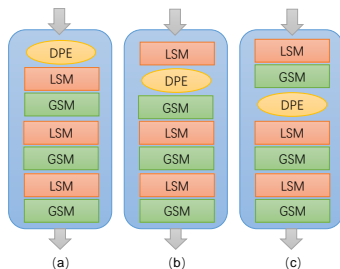
Method	Average	Aotra	Gallbladder	Kidnery (L)	Kidnery (R)	Liver	Pancreas	Spleen	Stomach
After the 1st LSM	87.91	91.60	77.14	<b>93.97</b>	90.31	95.64	<b>77.39</b>	90.64	<b>86.59</b>
After the 1st GSM	88.20	91.85	78.16	93.51	91.31	95.68	76.00	92.72	86.36
Before the 1st LSM (ours)	<b>88.83</b>	<b>92.12</b>	<b>80.09</b>	92.60	<b>91.91</b>	<b>96.99</b>	76.67	<b>93.78</b>	86.44

**Effect of global scope module (GSM).** To investigate the necessity of the Global Scope Module (GSM), we replace it by the Local Scope Module (LSM), with the other architectural components unchanged. As shown in Table 4, one can see that the GSM is beneficial to the segmentation accuracy, outperforming using only LSM modules by 0.66% in the average DSC. This verifies the necessity to explore global interactions of patches across units.

**Global scope multi-head self-attention (GS-MSA) vs. other self-attention.** In order to confirm the effectiveness of our GS-MSA, we compare it with the shift window strategy proposed in Swin Transformer [46] which achieves state-of-the-art performance in multiple computer vision tasks. Similar to our GS-MSA design, the shift window strategy (SW-MSA) aims to introduce global attention. Table 5 shows that our global attention design surpasses that in Swin Transformer by 1.33% in the average DSC.

**Dynamic position encoding vs. other position encodings.** We compare our dynamic position encoding (DPE) with other common position encoding methods, including the relative position encoding (RPE) [46, 69], absolute position encoding (APE) [34], and sinusoidal position encoding (SPE) [34]. The results are shown in Table 6. Compared to APE, SPE, and RPE, our DPE improves them by 4.05%, 2.79%, and 2.48% in the average DSC, respectively.

**Position of the dynamic position encoding block.** We conduct experiments to examine the performances of different choices of positions to apply the dynamic position encoding block, including placing it (a) before the first LSM, (b) right after the first LSM, and (c) right after the first GSM, in every D-Former block, as illustrated in Fig. 8 taking D-Former block 3 as an example. From Table 7, one may find that introducing the position information before the first LSM provides the best segmentation outcomes.

**Table 8:** Different positions to apply the DPE block. D-Former block 3 is used as an example for illustration, which contains three LSMs and three GSMs, arranging in an alternate manner.



## 5 Conclusions

In this paper, we proposed a novel 3D medical image segmentation framework called D-Former, which utilizes the common U-shape encoder-decoder design and is constructed based on our new dilated Transformer. Our proposed D-Former model can achieve both good efficiency and accuracy, due to its reduced number of patches used in self-attention in local scope module (LSM) and its exploration of long-range dependency with a dilated scope of attention in global scope module (GSM). Moreover, we introduced the dynamic position encoding block, making it possible to flexibly learn vital position information within input sequences. In this way, our model not only reduces the model parameters and decreases the FLOPs, but also attains state-of-the-art semantic segmentation performance on the Synapse and ACDC datasets.

**Acknowledgments.** This research was partially supported by the National Key R&D Program of China under grant No. 2019YFB1404802, National Natural Science Foundation of China under grants No. 62176231 and 62106218, Zhejiang public welfare technology research project under grant No. LGF20F020013, Wenzhou Bureau of Science and Technology of China under grant No. Y2020082. D. Z. Chen’s research was supported in part by NSF Grant CCF-1617735.

## References

- [1] Chen, J., Lu, Y., Yu, Q., et al.: TransUNet: Transformers make strong encoders for medical image segmentation. ArXiv Preprint ArXiv:2102.04306 (2021)
- [2] Peng, C., Zhang, X., Yu, G., *et al.*: Large kernel matters—Improve semantic segmentation by global convolutional network. In: CVPR (2017)
- [3] Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. ArXiv Preprint ArXiv:1511.07122 (2015)
- [4] Chen, L.-C., Papandreou, G., Kokkinos, I., et al.: DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. PAMI (2017)
- [5] Chen, L.-C., Zhu, Y., Papandreou, G., *et al.*: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: ECCV (2018)
- [6] Gu, Z., Cheng, J., Fu, H., et al.: CE-Net: Context encoder network for 2D medical image segmentation. TMI (2019)
- [7] Zhao, H., Shi, J., Qi, X., *et al.*: Pyramid scene parsing network. In: CVPR (2017)

- [8] Roth, H.R., Shen, C., Oda, H., *et al.*: A multi-scale pyramid of 3D fully convolutional networks for abdominal multi-organ segmentation. In: MICCAI (2018)
- [9] Feng, S., Zhao, H., Shi, F., *et al.*: CPFNet: Context pyramid fusion network for medical image segmentation. TMI (2020)
- [10] Dai, J., Qi, H., Xiong, Y., *et al.*: Deformable convolutional networks. In: ICCV (2017)
- [11] Heinrich, M.P., Oktay, O., Bouteldja, N.: OBELISK-Net: Fewer layers to solve 3D multi-organ segmentation with sparse deformable convolutions. MIA (2019)
- [12] Li, Z., Pan, H., Zhu, Y., *et al.*: PGD-UNet: A position-guided deformable network for simultaneous segmentation of organs and tumors. In: International Joint Conference on Neural Networks (2020)
- [13] Sun, Z., Cao, S., Yang, Y., *et al.*: Rethinking Transformer-based set prediction for object detection. In: ICCV (2021)
- [14] Pan, X., Xia, Z., Song, S., *et al.*: 3D object detection with pointformer. In: CVPR (2021)
- [15] Zhang, Z., Zhang, H., Zhao, L., *et al.*: Aggregating nested Transformers. ArXiv Preprint ArXiv:2105.12723 (2021)
- [16] Mehta, S., Koncel-Kedziorski, R., Rastegari, M., Hajishirzi, H.: DeFINE: DEep Factorized INput Token Embeddings for neural sequence modeling. ArXiv Preprint ArXiv:1911.12385 (2020)
- [17] Mehta, S., Ghazvininejad, M., Iyer, S., *et al.*: DeLighT: Very deep and light-weight Transformer. CoRR (2020)
- [18] Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR (2015)
- [19] Huang, H., Lin, L., Tong, R., *et al.*: UNet 3+: A full-scale connected UNet for medical image segmentation. In: IEEE International Conference on Acoustics, Speech and Signal Processing (2020)
- [20] Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: MICCAI (2015)
- [21] Mishra, P., Sarawadekar, K.: Polynomial learning rate policy with warm restart for deep neural network. In: IEEE Region 10 Conference (2019)
- [22] Yi-de, M., Qing, L., Zhi-Bai, Q.: Automated image segmentation using

- improved PCNN model based on cross-entropy. In: International Symposium on Intelligent Multimedia, Video and Speech Processing (2004)
- [23] Jadon, S.: A survey of loss functions for semantic segmentation. In: IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (2020)
- [24] Chu, X., Tian, Z., Zhang, B., et al.: Conditional positional encodings for vision Transformers. ArXiv Preprint ArXiv:2102.10882 (2021)
- [25] Kauderer-Abrams, E.: Quantifying translation-invariance in convolutional neural networks. ArXiv Preprint ArXiv:1801.01450 (2017)
- [26] Fu, S., Lu, Y., Wang, Y., et al.: Domain adaptive relational reasoning for 3D multi-organ segmentation. In: MICCAI (2020)
- [27] Wang, W., Chen, C., Ding, M., et al.: TransBTS: Multimodal brain tumor segmentation using Transformer. In: MICCAI (2021)
- [28] Xu, G., Wu, X., Zhang, X., et al.: LeViT-UNet: Make faster encoders with Transformer for medical image segmentation. ArXiv Preprint ArXiv:2107.08623 (2021)
- [29] Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., et al.: UNet++: A nested U-Net architecture for medical image segmentation. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, (2018)
- [30] Ni, Z.-L., Bian, G.-B., Zhou, X.-H., et al.: RAUNet: Residual attention u-net for semantic segmentation of cataract surgical instruments. In: International Conference on Neural Information Processing (2019)
- [31] Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., et al.: 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In: MICCAI (2016)
- [32] Isensee, F., Jaeger, P.F., Kohl, S.A., et al.: nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. Nature Methods (2021)
- [33] Diakogiannis, F.I., Waldner, F., Caccetta, P., et al.: ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data. Journal of Photogrammetry and Remote Sensing (2020)
- [34] Vaswani, A., Shazeer, N., Parmar, N., et al.: Attention is all you need. In: NIPS (2017)
- [35] Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al.: An image is worth

- 16x16 words: Transformers for image recognition at scale. ArXiv Preprint ArXiv:2010.11929 (2020)
- [36] Devlin, J., Chang, M.-W., Lee, K., et al.: Bert: Pre-training of deep bidirectional Transformers for language understanding. ArXiv Preprint ArXiv:1810.04805 (2018)
- [37] Carion, N., Massa, F., Synnaeve, G., et al.: End-to-end object detection with Transformers. In: ECCV (2020)
- [38] Zheng, S., Lu, J., Zhao, H., et al.: Rethinking semantic segmentation from a sequence-to-sequence perspective with Transformers. In: CVPR (2021)
- [39] Zhu, X., Su, W., Lu, L., et al.: Deformable DETR: Deformable Transformers for end-to-end object detection. ArXiv Preprint ArXiv:2010.04159 (2020)
- [40] Touvron, H., Cord, M., Douze, M., et al.: Training data-efficient image Transformers & distillation through attention. In: ICML (2021)
- [41] Wang, X., Girshick, R., Gupta, A., et al.: Non-local neural networks. In: CVPR (2018)
- [42] Zhou, H.-Y., Guo, J., Zhang, Y., et al.: nnFormer: Interleaved Transformers for volumetric segmentation. ArXiv Preprint ArXiv:2109.03201 (2021)
- [43] Huang, X., Deng, Z., Li, D., et al.: MISSFormer: An effective medical image segmentation Transformer. ArXiv Preprint ArXiv:2109.07162 (2021)
- [44] Lin, A., Chen, B., Xu, J., et al.: DS-TransUNet: Dual swin Transformer U-Net for medical image segmentation. ArXiv Preprint ArXiv:2106.06716 (2021)
- [45] Wang, W., Xie, E., Li, X., et al.: Pyramid vision Transformers: A versatile backbone for dense prediction without convolutions. ArXiv Preprint ArXiv:2102.12122 (2021)
- [46] Liu, Z., Lin, Y., Cao, Y., et al.: Swin Transformers: Hierarchical vision Transformers using shifted windows. ArXiv Preprint ArXiv:2103.14030 (2021)
- [47] Yuan, L., Chen, Y., Wang, T., et al.: Tokens-to-Token ViT: Training vision Transformers from scratch on ImageNet. ArXiv Preprint ArXiv:2101.11986 (2021)
- [48] Yuan, L., Hou, Q., Jiang, Z., et al.: VOLO: Vision outlooker for visual

- recognition. ArXiv Preprint ArXiv:2106.13112 (2021)
- [49] Hatamizadeh, A., Tang, Y., Nath, V., et al.: UNETR: Transformers for 3D medical image segmentation. ArXiv Preprint ArXiv:2103.10504 (2021)
  - [50] Xie, Y., Zhang, J., Shen, C., et al.: CoTr: Efficiently bridging CNN and Transformer for 3D medical image segmentation. ArXiv Preprint ArXiv:2103.03024 (2021)
  - [51] Zhang, Y., Liu, H., Hu, Q.: TransFuse: Fusing Transformers and CNNs for medical image segmentation. ArXiv Preprint ArXiv:2102.08005 (2021)
  - [52] El-Nouby, A., Touvron, H., Caron, M., et al.: XCiT: Cross-covariance image Transformers. ArXiv Preprint ArXiv:2106.09681 (2021)
  - [53] Cao, H., Wang, Y., Chen, J., et al.: Swin-Unet: Unet-like pure Transformer for medical image segmentation. ArXiv Preprint ArXiv:2105.05537 (2021)
  - [54] Wu, Z., Liu, Z., et al.: Lite Transformer with long-short range attention. ArXiv Preprint ArXiv:2004.11886 (2020)
  - [55] Korez, R., Likar, B., Pernuš, F., *et al.*: Model-based segmentation of vertebral bodies from MR images with 3D CNNs. In: MICCAI (2016)
  - [56] Zhou, X., Ito, T., Takayama, R., *et al.*: Three-dimensional CT image segmentation by combining 2D fully convolutional network with 3D majority voting. In: Deep Learning and Data Labeling for Medical Applications, (2016)
  - [57] Moeskops, P., Wolterink, J.M., *et al.*: Deep learning for multi-task medical image segmentation in multiple modalities. In: MICCAI (2016)
  - [58] Shakeri, M., Tsogkas, S., Ferrante, E., *et al.*: Sub-cortical brain structure segmentation using F-CNN's. In: International Symposium on Biomedical Imaging (2016)
  - [59] Alansary, A., Kamnitsas, K., Davidson, A., *et al.*: Fast fully automatic segmentation of the human placenta from motion corrupted MRI. In: MICCAI (2016)
  - [60] Wang, C., MacGillivray, T., Macnaught, G., et al.: A two-stage 3D Unet framework for multi-class segmentation on full resolution image. ArXiv Preprint ArXiv:1804.04341 (2018)
  - [61] Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., *et al.*: 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In: MICCAI (2016)

- [62] Kamnitsas, K., Ledig, C., Newcombe, V.F., et al.: Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. MIA (2017)
- [63] Drozdal, M., Vorontsov, E., Chartrand, G., *et al.*: The importance of skip connections in biomedical image segmentation. In: Deep Learning and Data Labeling for Medical Applications, (2016)
- [64] Ghafoorian, M., Karssemeijer, N., Heskes, T., *et al.*: Non-uniform patch sampling with deep convolutional neural networks for white matter hyperintensity segmentation. In: International Symposium on Biomedical Imaging (2016)
- [65] Brosch, T., Tang, L.Y., Yoo, Y., et al.: Deep 3D convolutional encoder networks with shortcuts for multiscale feature integration applied to multiple sclerosis lesion segmentation. TMI (2016)
- [66] Christ, P.F., Ettliger, F., et al.: Automatic liver and tumor segmentation of CT and MRI volumes using cascaded fully convolutional neural networks. ArXiv Preprint ArXiv:1702.05970 (2017)
- [67] Brosch, T., Tang, L.Y., Yoo, Y., et al.: Deep 3D convolutional encoder networks with shortcuts for multiscale feature integration applied to multiple sclerosis lesion segmentation. TMI (2016)
- [68] Pereira, S., Pinto, A., et al.: Brain tumor segmentation using convolutional neural networks in MRI images. TMI (2016)
- [69] Shaw, P., Uszkoreit, J., Vaswani, A.: Self-attention with relative position representations. ArXiv Preprint ArXiv:1803.02155 (2018)
- [70] Cai, S., Tian, Y., Lui, H., et al.: Dense-UNet: A novel multiphoton in vivo cellular image segmentation model based on a convolutional neural network. Quantitative Imaging in Medicine and Surgery (2020)
- [71] Han, K., Xiao, A., Wu, E., et al.: Transformer in Transformer. ArXiv Preprint ArXiv:2103.00112 (2021)
- [72] Valanarasu, J.M.J., Oza, P., et al.: Medical Transformer: Gated axial-attention for medical image segmentation. ArXiv Preprint ArXiv:2102.10662 (2021)
- [73] Chen, L.-C., Papandreou, G., Kokkinos, I., et al.: Semantic image segmentation with deep convolutional nets and fully connected CRFs. ArXiv Preprint ArXiv:1412.7062 (2014)
- [74] Chen, L.-C., Papandreou, G., Kokkinos, I., et al.: DeepLab: Semantic

image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. TPAMI (2017)

- [75] Chen, L.-C., Papandreou, G., Schroff, F., et al.: Rethinking atrous convolution for semantic image segmentation. ArXiv Preprint ArXiv:1706.05587 (2017)
- [76] Chen, L.-C., Zhu, Y., Papandreou, G., *et al.*: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: ECCV (2018)
- [77] Milletari, F., Navab, N., Ahmadi, S.-A.: V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In: 3DV (2016)
- [78] Schlemper, J., Oktay, O., Schaap, M., et al.: Attention gated networks: Learning to leverage salient regions in medical images. MIA (2019)
- [79] Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. ArXiv Preprint ArXiv:1607.06450 (2016)
- [80] Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In: CVPR, pp. 1251–1258 (2017)
- [81] Deng, J., Dong, W., Socher, R., *et al.*: ImageNet: A large-scale hierarchical image database. In: CVPR (2009)
- [82] Bottou, L.: Stochastic gradient descent tricks. In: Neural Networks: Tricks of the Trade, (2012)