M.S. THESIS

Multi-Organ Segmentation for CT Images through Deformable Window Attention

비정형 윈도우 어텐션을 이용한 CT에서의 다중 장기 분할

2023 년 2 월

서울대학교 대학원

컴퓨터 공학부

김 재 용

Multi-Organ Segmentation for CT Images through Deformable Window Attention

비정형 윈도우 어텐션을 이용한 CT에서의 다중 장기 분할

지도교수 신 영 길

이 논문을 공학석사학위논문으로 제출함

2022 년 11 월

서울대학교 대학원

컴퓨터 공학부

김 재 용

김재용의 공학석사 학위논문을 인준함

2022 년 12 월

위 원 장	서 진 욱	(인)
부위원장	신 영 길	(인)
위 원	이 영 기	(인)

Abstract

Multi-organ segmentation is a crucial task for clinical applications of computer-aided diagnosis. Recent development in deep learning, especially convolutional neural networks(CNN), showed promising outcomes on the simultaneous segmentation of multiple organs in medical images. However, most approaches with a backbone consisting of CNNs tend to weakly relate to global feature representations due to the limited shape of convolutions. As architecture modifications are being explored, vision transformers(ViT) have been displaying significant improvements even surpassing the performance of CNNs on image classification tasks. In this paper, to increase the performance of multi-organ segmentation, inspired by deformable convolutional networks, we introduce a deformable attention network that learns offsets and scales in order to focus the attention to more informative areas of the image rather than focusing on comparing each patch to the whole image. The overall architecture effectively utilizes both CNNs and ViTs to not only increase accuracy but also reduce computational complexity of self-attention mechanisms. We used the Beyond the Cranium Vault(BTCV) dataset which contains only 30 CT abdominal images for training and validation, and 17 CT abdominal images for testing. The experimental results show that the proposed network produces more accurate results compared to previous methods by achieving a 3.7% increase in pancreas segmentation and a 3.6% increase in duodenum segmentation in terms of dice similarity coefficient (DSC) score.

Keywords: deep neural network, deformable attention, multi-organ segmentation, biomedical image segmentation

Student Number: 2020-25153

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Multi-class segmentation of organs in abdominal computed tomography (CT) scans is an essential task for clinical applications of computer-aided diagnosis (CAD) [1, 2]. It also plays an integral role in medical image analysis as it is often the first step for analysis of anatomical structures. Thus, it is critical to acquire accurate and reliable segmentation results to optimize clinical workflow. However, due to the unclear boundaries and variability in shapes of organs (illustrated in Figure 1.1) make it cumbersome even for medical experts to accurately segment organs. Furthermore, manual or semi-automatic segmentation could cause potential fatigue of human experts. In order to overcome such problems, machine learning techniques have been actively studied.

Since the advent of deep learning in medical image segmentation, convolution neural networks(CNN) and especially fully convolutional neural networks (FCNNS) [3] have become dominant in medical image segmentation tasks. In particular, U-Net [4], which consists of a U-shaped symmetric encoder-decoder architecture with skip connections to enhance detail retention, have inspired

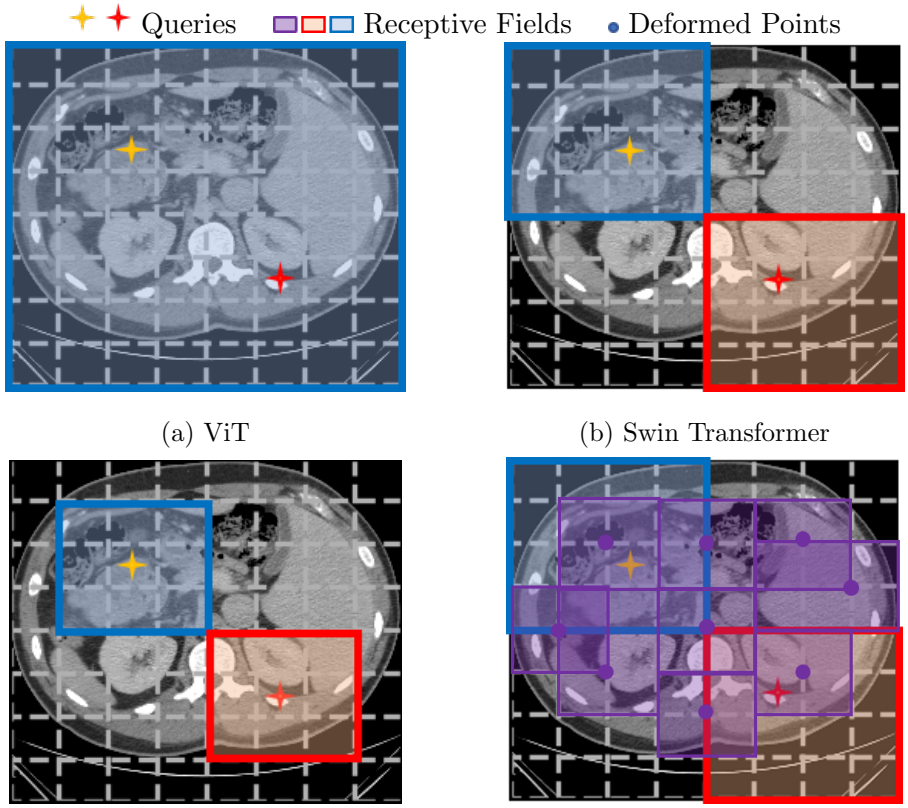
Figure 1.1: Abdominal CT image, manual multi-organ segmentation image, and overlapped image

many works that have achieved state-of-the art results in various medical semantic segmentation tasks. Although such approaches give the network powerful representation learning capabilities, there are limitations when it comes to learning long-ranged dependencies due to their localized receptive fields [5, 6]. Ultimately, their incapability to learn such long-ranged dependencies leads to sub-optimal results in segmentation of structures with various shapes and sizes. Various architectural modifications have been suggested by researchers for an efficient solution over the course of time and this leads to attention mechanisms.

Transformers have been showing state-of-the-art performance in natural language processing(NLP) tasks. The self-attention mechanism in transformers highlights the important features of word sequences [7, 8]. Only recently, transformers have been applied to computer vision tasks. Unlike convolutions, transformers treat images as a sequence of 1D patch embeddings. Utilizing self-attention modules, transformers are able to learn the relations between patches in a global manner, thus effectively learns long-range dependencies. However, there are problems that transformers may not capture localized spatial features as well as convolutions and furthermore need a large dataset to outperform similar-sized CNN counterparts. Also, when it comes to high-resolution images

for dense predictions such as segmentation, a global self-attention may become burdensome due to the quadratic computational cost with respect to the number of grids in feature maps.

In this study, we aim at multi-organ segmentation in CT images using transformers and CNNs. To resolve the aforementioned disadvantages of CNNs and transformers, we propose a model that effectively reduces computational cost and strengthens the ability of transformers to encode important local structure simultaneously. Our overall model translates the task of 3D segmentation into a 1D sequence-to-sequence prediction problem. It uses pure transformers as encoders instead of CNN-based encoders for learning long-ranged dependencies, and features extracted from the encoder are merged with the CNN-based decoder via skip-connection to complement local spatial information to produce a segmentation output. Furthermore, we implement SwinTransformer[9] as the general backbone encoder along with a deformable attention module inspired by [10, 11]. Unlike the original vision transformer, our model constructs a hierarchical representation by starting from small-sized patches and gradually merging neighboring patches in deeper transformer layers. The main difference between our model and the previous works is that, along in the process of building a hierarchical representation, the deformable attention module enhances the model's capability of learning more informative regions. Considering that learning a deformable receptive field for the convolution filters has been shown effective in a data dependent setting, we attempt to corporate deformable attention patterns into our encoder. As illustrated in Figure 1.2(d), it does so by adding 3D offsets to the regular grid sampling locations within the local attention window. The offsets are learned to shift keys and values in self-attention to important regions. This design reduces the quadratic cost to a linear space complexity and introduces the self-attention module to more informative features.



(a) ViT (b) Swin Transformer

(c) CNN (d) Ours

Figure 1.2: Comparison of our proposed method with other transformer models and CNN. The yellow and red star indicate each queries, and the masks with solid boundaries show the receptive fields for each query according to the model. (a) ViT adopts full attention for all queries. (b) Swin Transformer adopts partitioned window attention. (c) CNN uses convolutional filters to extract features. (d) Our model uses window-partitioning along with learning deformed points for all queries

Chapter 2

Related Works

2.1 Image Segmentation

Image Segmentation is the process of partitioning an image into image regions or a set of pixels. The result of a segmentation usually consists of a set of segments that collectively covers the entire image, or a set of contours extracted from the image. As pixels contain computed properties, such as color and intensity, many efforts have been made to classify pixels to regions, that are characterized by similar values of color or intensity, using methods ranging from region growing methods to graph partitioning methods.

Region growing methods [12] are used based on the assumption that the neighboring pixels within a region have similar values. It compares one pixel with its neighbors, and if the similarity criterion is satisfied, the pixel will become a part of the same cluster as one or more of its neighbors. However, the results of region growing based methods are significantly influenced by noise in images and difficult to apply in real life images because all images practically

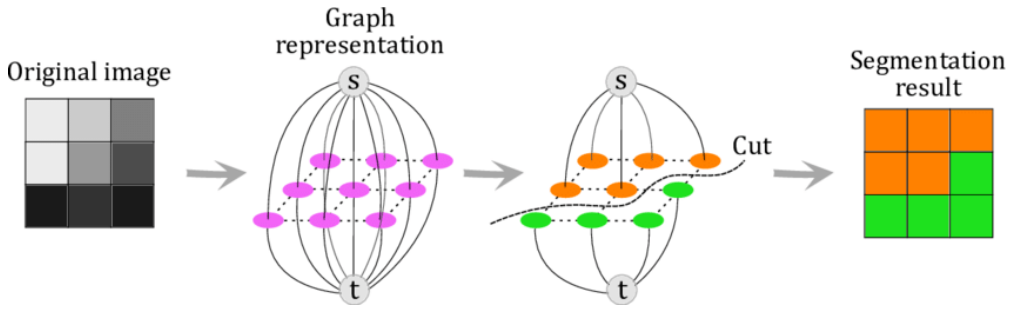


Figure 2.1: Diagram of graph cut method in segmentation

have noise. Especially in the context of medical image segmentation, due to organs having unclear boundaries, such methods are not so robust.

More sophisticated attempts were made by using graph cuts [13]. Apart from the previous methods that optimized functions defined on a continuous contour or surface, graph cut optimizes a cost function defined on a discrete set of variables. It uses a cost function, that can include both region and boundary properties of segments, to determine whether the pixel in inside or outside the object of interest. As shown in Figure 2.1, a network flow is built based on an undirected graph with two terminal nodes S and T that represents object and background labels respectively. There are two type of edges, called n-link and t-links. Both links each carry a weight, in which are optimized in order to sever the correct edges for an optimal segmentation. Graph cuts methods became popular for optimizing the location of a contour, but there exists problems where the memory usage increases quickly as the image size increases and maybe unfit for classifying pixels for multiple labels as it is only able to find a global optimum for binary labeling, such as foreground/background image segmentation.

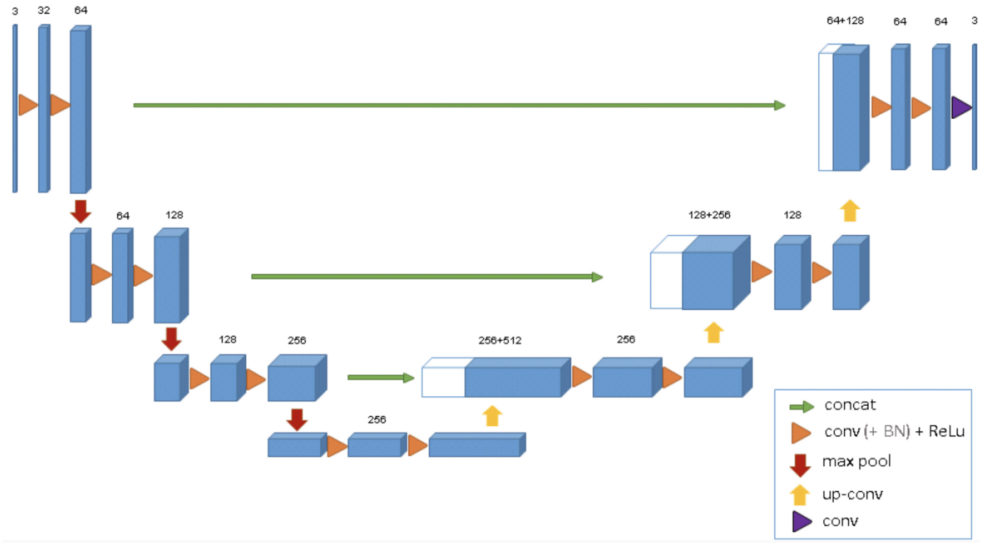


Figure 2.2: UNet architecture

2.2 CNN-based Segmentation Networks

Since the introduction of the U-Net [4], CNN-based networks have shown groundbreaking performance on various 2D and 3D medical image segmentation tasks. As shown in Figure 2.2, the method uses the encoder-decoder based architecture, using 3D convolutional networks in order to capture rich features in various resolutions, along with skip connections to generate a fine-level dense prediction in their original resolutions. Many of the following works were inspired by this form of encoder-decoder networks for medical image segmentation [14, 15]. Furthermore, there were studies that complement such networks by applying contour aware modules [16] and shape-aware modules [17]. By taking advantage of essential spatial information such as contours and shape information, such endeavors showed improvement worth noticing and shows that the U-Net architecture is also flexible as a segmentation backbone. These methods

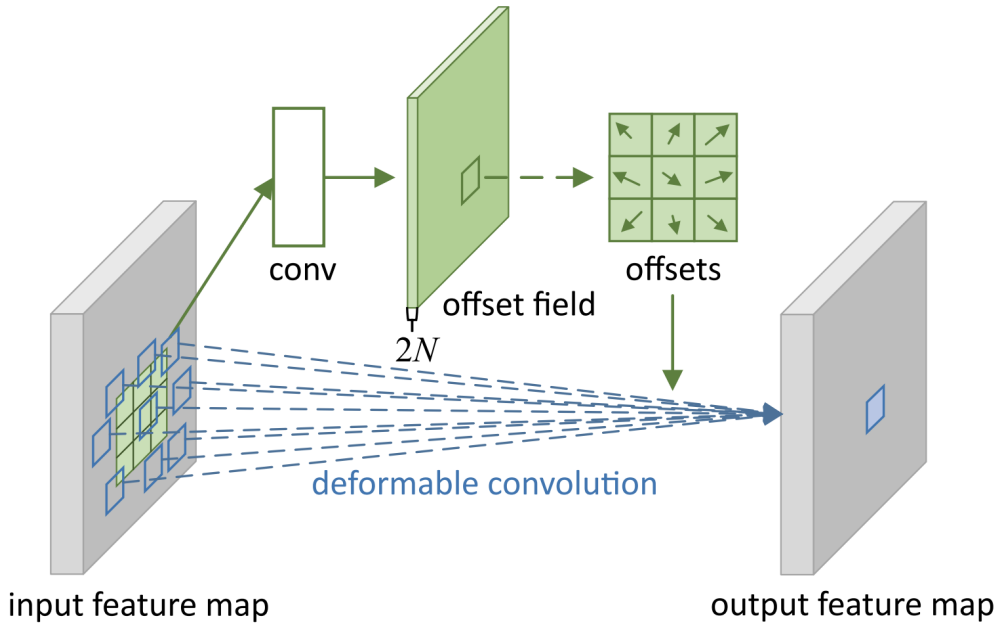


Figure 2.3: Illustration of 3×3 deformable convolution

effectively reduced problems in spatial context and low-resolution condition. However, there is a problem that these networks have difficulties in learning global context and long-range spatial dependencies, which could be crucial to the segmentation performance.

2.3 Vision Transformers

Transformers were first proposed by Vaswani et. al [8] and were mainly used in machine translation and achieved state-of-the-art in many natural language processing tasks. Transformers have recently gained attention for computer vision tasks as the vision transformers were introduced by Dosovitskiy et al. [18] Specifically, images were interpreted as a sequence of patches and processed by a standard transformer encoder as it is used in NLP. Illustrated in Fig-

ure 2.5, a transformer encoder usually consists of a multi-head self-attention layer(MHSA), shown in Figure 2.4, and a MLP block. An attention function can be described as mapping a query and a set of key-value pairs to an output, where query, keys, and values, and output are all vectors. Basically, the outputs are values that indicate the relation between the query and the corresponding key. Recently, Vision transformer(ViT) demonstrated state-of-the-art performance on image classification datasets by employing transformers with global self-attention to images.

Further studies, on using ViTs as backbones, focused on reducing the complexity of self-attention blocks. Because each query is compared to the whole image, the computation cost is quadratic depending on the size of the input image and patch size. Liu et al. [9] proposed Swin-Transformer, which extracts feature representations at several resolutions with a shifted windowing mechanism for computing the self-attention. Linear computational complexity is achieved by computing self-attention locally within non-overlapping windows with fixed number of patches in each local window. Attempts to apply in the context of medical image segmentation had been made [19], but were limited to 2D inputs where 3D volumetric images were cut into slices in order to infer.

Recently, multiple methods were proposed that exploit both transformers and CNNs. Chen at el. [20] proposed a hybrid CNN-transformer architecture where a transformer is applied as an additional layer in the bottleneck of a U-Net architecture. The proposed idea was able to leverage both global contexts encoded by transformers and detailed high-resolution local spatial information from CNN features, but uses slices of volumetric images. Hatamizadeh et al. [21] introduced a model that uses pure transformers as encoders and CNNs as decoders, but the computationally inefficient and needs a larger dataset to perform better than CNN-based models. To this end, our method produces im-

proved results by: (1) Taking full three-dimensional volumes as inputs instead of taking two-dimensional slices from volumes and later restore the volume through post-processing. (2) Utilizing Swin Transformer and a deformable attention module for reducing computational complexity, and focusing attention to more informative features.

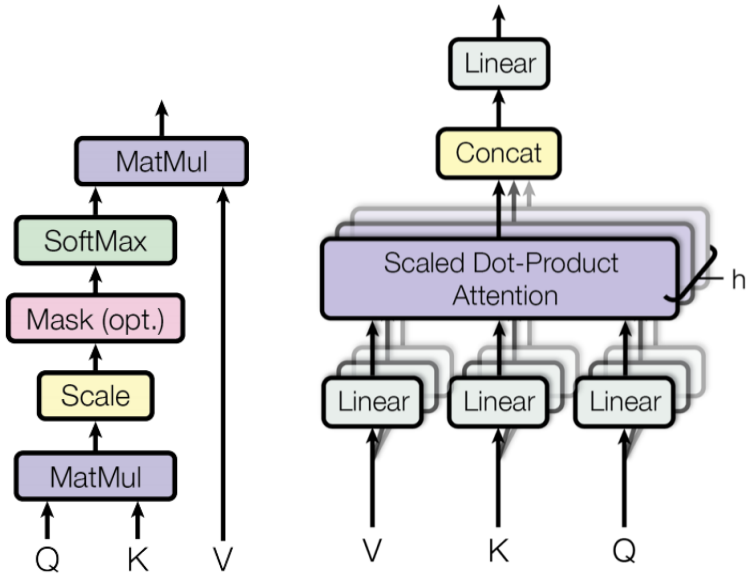


Figure 2.4: Multi-head self attention(MHSA) block

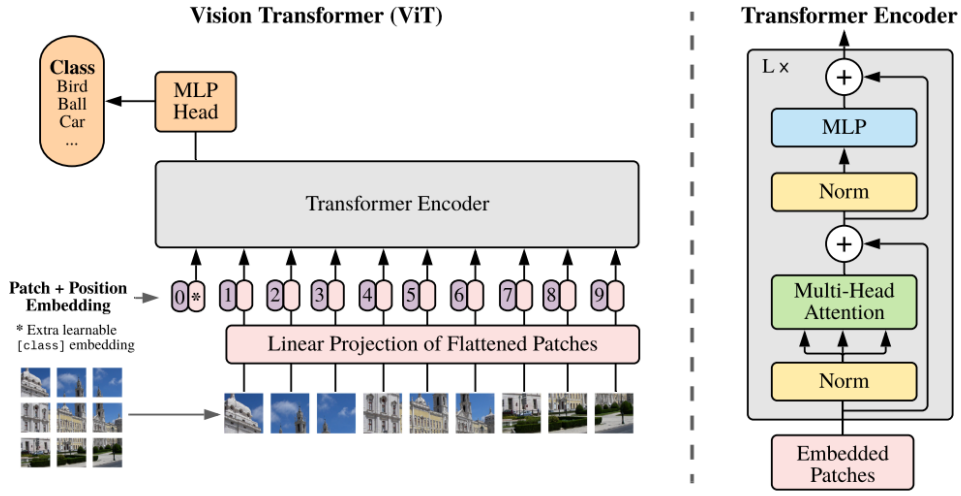


Figure 2.5: Vision Transformer(ViT) architecture

Chapter 3

Methodology

3.1 Overview

In this chapter, the proposed architecture for multi-organ segmentation is introduced. As mentioned in chapter 3, we present a transformer-CNN hybrid model for the abdominal organ segmentation based on Swin Transformer and our proposed deformable attention module. The overall architecture will be explained first, and then specific building blocks will be explained.

3.2 Overall Architecture

Our proposed architecture is shown in Figure 3.1. Overall, the architecture has a encoder-decoder scheme resembling 3D-UNet [4] except that the encoder is developed with pure transformers. First, the input to the model $x \in \mathcal{R}^{H \times W \times D \times S}$, where H, W, D, are height, width, and depth, is a token with a patch resolution of $(H^{'}, W^{'}, D^{'})$. In order to tokenize the image into patches, a patch partition layer is utilized. The tokens are projected into an embedding space with di-

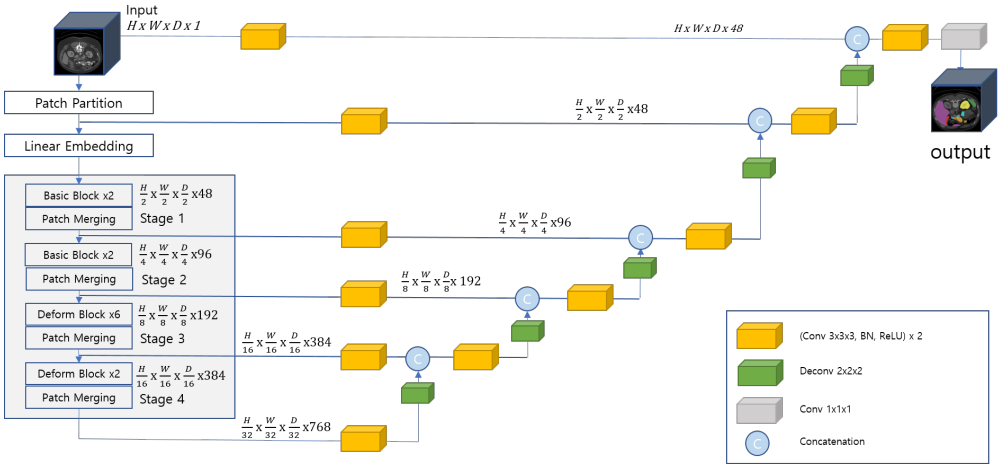


Figure 3.1: Overview of our proposed architecture

mension C and then proceeds to the backbone of the network. The backbone encoder of our model takes a shape of a pyramid in order to capture features in a hierarchical fashion. It consists of 4 stages with 2 basic Swin Transformer [9] blocks and 2 deformable transformer blocks in the latter two stages. After each stage is completed, a patch merging layer is utilized to decrease the resolution of feature representations by 2 in order to maintain the hierarchical structure. The feature representations from the encoder are fed to a residual block, that consists of two 3D convolutional layers that are normalized by instance normalization and activated by ReLU, via skip-connection. Subsequently, the resolution of the feature maps are increased by a factor of 2 when going through a deconvolutional layer and the outputs are concatenated with the outputs of the previous stage. The concatenated features are again fed into another residual block as described. The final segmentation outputs are computed by using a $1 \times 1 \times 1$ convolutional layer and a sigmoid activation function.

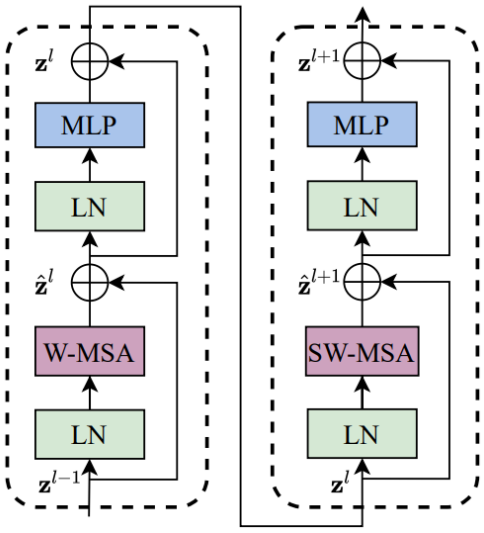


Figure 3.2: Diagram of the two successive basic block in our proposed network

3.3 Basic Transformer Block

This section highlights the "basic block" shown in Figure 3.1 and the detailed diagram is shown in Figure 3.2. At a given layer l in the transformer encoder, M is the size of windows to evenly partition a 3D token into $\lceil \frac{H'}{M} X \frac{W'}{M} X \frac{D'}{M} \rceil$, where (H', W', D') are patch resolutions in height, width, and depth respectively. Subsequently, in layer l+1, the partitioned window regions are shifted by $\lfloor \frac{M}{2} \rfloor, \lfloor \frac{M}{2} \rfloor, \lfloor \frac{M}{2} \rfloor$ voxels. Therefore, the outputs of the two successive basic blocks are calculated as

$$\begin{aligned} \hat{z}^l &= WMSA(LN(\hat{z}^{l-1})) + z^{l-1} \\ z^l &= MLP(LN(\hat{z}^l)) + \hat{z}^l \\ \hat{z}^{l+1} &= SWMSA(LN(\hat{z}^l)) + z^l \\ z^{l+1} &= MLP(LN(\hat{z}^{l+1})) + \hat{z}^{l+1} \end{aligned} \tag{3.1}$$

W-MSA and SW-MSA are regular and window partitioning multi-head self-attention modules. z^l and z^{l+1} denote the outputs of W-MSA and SW-MSA respectively. MLP and LN denote Multi-Layer Perception and Layer Normalization respectively. For efficient computation of partitioned and shifted windows, 3D cyclic-shifting [9] is leveraged. Self-attention in this case is computed as

$$Attention(Q, K, V) = Softmax(\frac{QK^T}{\sqrt{d}} + B)V \tag{3.2}$$

where $Q, K, V \in \mathcal{R}^{M^3 \times d}$ are the query, key, and value matrices. d is the query/key dimension, and M^2 is the number of patches in a window. $B \in \mathcal{R}^{M^3 \times M^3}$ is the relative position bias that is placed in each head in computing similarity.

3.4 Deformable Attention Layer

The "deform block" has a similar structure as the basic block, but instead of the shifted-windows self-attention module, the novel deformable attention module is applied. As illustrated in Figure , given the input feature map $x \in \mathcal{R}^{H \times W \times D \times C}$, a uniform grid of points $p \in \mathcal{R}^{H_G \times W_G \times D_G \times 3}$ are generated as reference points. When generating the uniform grid, the grid size is downsample from the input feature map size by a factor r, $H_G = H/r, W_G = W/r, D_G = D/r$, due to the computational expenses of learning the offsets for all the reference points. The values of reference points are linearly spaced 3D coordinates, and then we normalize them to the range [-1, +1] for every axis. In order to obtain the offsets for each reference point, the feature maps are projected linearly as tokens then fed to a small offset network $\varphi_{offset}(\cdot)$ shown in Figure 3.4. It consists of a 3D convolutional layer with a given stride, a GeLU activation layer, and finally a $1 \times 1 \times 1$ convolutional layer to obtain the normalized x, y, and z values along

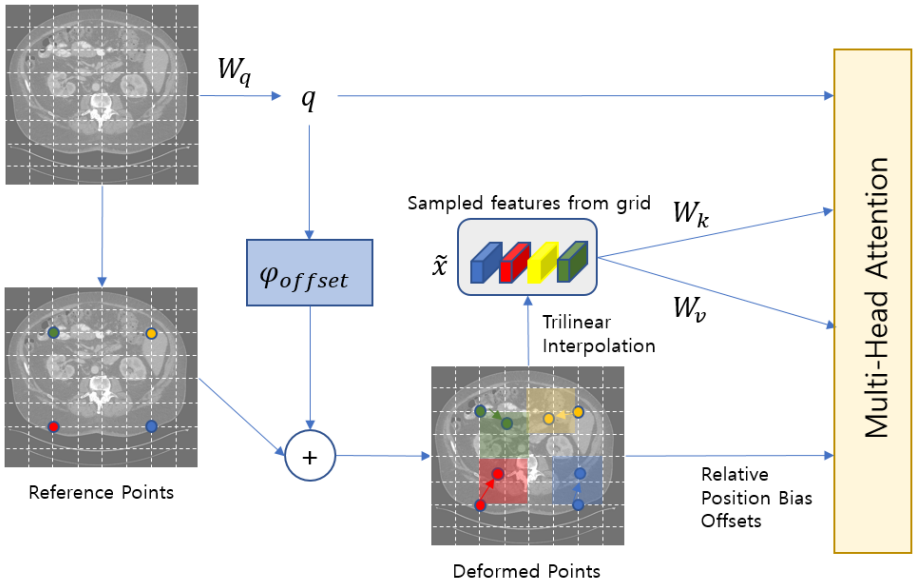


Figure 3.3: Diagram of the Deformable Attention Module

with the receptive window size. The output values from the offset network are added to the reference points and finally features are sampled at the locations of deformed points as keys and values:

$$q = xW_q, \hat{k} = \hat{x}W_k, \hat{v} = \hat{x}W_v \tag{3.3}$$

$$with \Delta p = \varphi_{offset}(q), \hat{x} = \Phi(x; p + \Delta p) \tag{3.4}$$

where $\hat{k}$ and $\hat{v}$ denote deformed key and value embeddings respectively. $\Phi(;)$ is the sampling function, which in this case is a trilinear interpolation to make it differentiable. We perform multi-head attention on q, k, v along with the relative position offsets R. The output of the self-attention head could be formulated as:

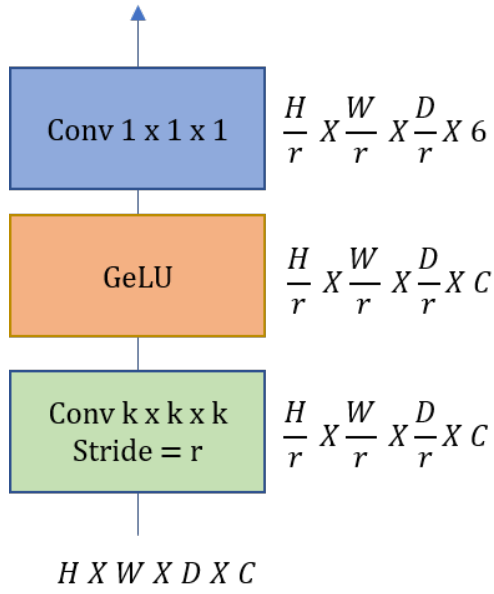


Figure 3.4: Diagram of the offset network

$$Attention(Q, K, V) = Softmax(\frac{QK^T}{\sqrt{d}} + \Phi(R))V \tag{3.5}$$

As already mentioned, offset values for reference points are generated by a sub-network, which consumes the query features. Taking in account that each reference point covers a local $s \times s$ region (s is the maximum value for offset), the generation network should also be able to pick up local spatial features to learn reasonable offsets. Therefore, the sub-network contains two convolution modules with a nonlinear activation. The kernel size and stride is up to the user but in our case, the best results showed when s=4, for both stages and kernel size (8, 8, 8) when stage 3 and (6, 6, 6) when stage 4.

The deformable attention module has a similar computation cost as the Swin Transformer. The only addition overhead is the offset network that generates offsets. The complexity of the whole module can be summarized as:

$$\Omega(DAM) = 2HWDN_sC + 2HWDC^2 + 2N_sC^2 + (k^3 + 3)N_sC \tag{3.6}$$

where the prior part of the equation is the complexity for the vanilla 3D self-attention module and the latter is the complexity for the offset network. $N_s = H_G W_G D_G = HWD/r^3$ is the number of sampled points in the equation. It is obvious from the equation that the cost of the offset network has linear complexity. Additionally, by choosing a larger downsampling factor r, the complexity will decrease drastically.

3.5 Overall Loss Function

As data imbalance problem between foreground and background is probable, we used a mixture of dice loss [22] and cross-entropy loss [23] to formulate our loss function. It is computed in a voxel-wise manner as

$$\mathcal{L}_{DL}(G, Y) = 1 - 2\sum_{j=1}^{J} \frac{\sum_{i=1}^{I} G_{i,j} Y_{i,j}}{\sum_{i=1}^{I} G_{i,j}^2 + \sum_{i=1}^{I} Y_{i,j}^2} \tag{3.7}$$

$$\mathcal{L}_{CE}(G, Y) = \sum_{j=1}^{J} Y_j \log(G_j) \tag{3.8}$$

$$\mathcal{L} = \lambda_1 \mathcal{L}_{DL} + \lambda_2 \mathcal{L}_{CE} \tag{3.9}$$

where I denotes voxel numbers; J is the number of classes; $Y_{i,j}$ and $G_{i,j}$ represent the probability of output and one-hot encoded ground truth for class j at voxel i; λ indicates the weight of each loss term respectively. In this study, we set both weights to 1 in the experiments.

Chapter 4

Experiment Details

4.1 Dataset

For the dataset, we utilized the Beyond the Carnial Vault (BTCV) dataset [24] for segmentation in CT imaging modalities. The BTCV dataset consists of 30 abdominal CT images for training and 17 CT images for testing. From the dataset, we referenced standard segmentations of the spleen, left kidney, gallbladder, esophagus, liver, stomach, pancreas, and duodenum. Each CT image contains 80 to 225 slices of 512×512 pixel images, where pixel sizes is in the range of 0.6 to 1.0mm, and slice thickness ranging from 1 to 6 mm. Among the 30 CT images, 24 images were used for training and 6 images were used for validation.

Each CT volumes were pre-processed independently by resampling and intensity normalization. All images were resampled into the isotropic voxel spacing of 1.5mm x 1.5mm x 2.0mm. After resampling, backgrounds were cropped as much as possible and the intensities were normalized to [0, 1] from the range

of [-175, 250] Hounsfield Units (HU). Finally, we randomly cropped the input images so that the input resolution is $128 \times 128 \times 128$ during training.

4.2 Implementation Details

Embed Dimension	Feature Size	Number of Blocks	Window Size	Number of Heads
768	48	[2, 2, 6, 2]	[7, 7, 7]	[3, 6, 12, 24]

Table 4.1: Our model configurations

Our transformer based encoder and CNN based decoder takes a 3D U-Net architecture as the base network. Aiming to build a hierarchical feature pyramid, the encoder backbone includes 4 stages with patch merging. Table 4.1 elaborates the details of the configurations of our architecture. For our deformable attention module, in stage 3, the downsampling rate for the grid was 4, offset scale was set to 4, and the kernel size for the offset convolution was set to (8, 8, 8). In stage 4, the downsampling rate for the grid was 2, offset scale was set to 4, and the kernel size for the offset convolution was set to (6, 6, 6). In addition to the aforementioned pre-processing, we used data augmentation methods, such as random flip in axial, sagittal, and coronal views, random rotation of 90, 180, 270 degrees, and random intensity shift in the range of [-0.1, 0.1]. We used a batch size of 1, the AdamW optimizer [25] with an initial learning rate of 0.0001 for 25,000 iterations.

For inference, we used a sliding window approach with an overlap potion of 0.8 between the neighboring patches. The proposed method is implemented through PyTorch and MONAI, an open-source python library for medical image processing. The model was trained using a NVIDIA TITAN XP 12GB.

Method	DSC								
	avg	**spleen**	**left_kidney**	**gallbladder**	**esophagus**	**liver**	**stomach**	**pancreas**	**duodenum**
3D-UNet [4]	0.721	0.902	0.889	0.635	0.600	0.919	0.764	0.740	0.591
V-Net [14]	0.719	0.888	0.862	0.612	0.525	0.912	0.721	0.698	0.535
Attention-UNet [15]	0.703	0.922	0.530	0.589	0.524	0.949	0.700	0.7613	0.6474
TransUNet [20]	0.723	0.911	0.819	0.629	0.595	0.941	0.756	0.569	0.548
UNETR [21]	0.750	0.916	0.833	0.673	0.579	0.949	0.795	0.693	0.561
SWIN-UNETR [26]	0.796	0.942	**0.932**	0.687	**0.608**	0.953	0.849	0.752	0.645
Ours	**0.813**	**0.950**	0.931	**0.727**	0.604	**0.959**	**0.860**	**0.789**	**0.681**

Table 4.2: Dice Similarity Coefficient score of our proposed network and previous methods

4.3 Results

Quantitative Results Table 4.2 shows the quantitative results of multi-organ segmentation. It shows the Dice score Coefficients for each organ. We evaluated the performance of our model with other models that use only CNNs and models that use both transformers and CNNs. Compared models are 3D-UNet, V-Net, TransUNet. For a fair comparison, we did not perform any post-processing for the output on any of the models. For our evaluation metric, we used Dice score coefficient (DSC) [22]. The results show that our proposed model outperforms other models in most organs even for smaller organs such as the pancreas and duodenum by approximately 0.037 and 0.046 respectively.

Qualitative Results The qualitative results are illustrated in Figure 4.1. Input images and results are shown in axial slices. Results show that our model outperforms other networks. Notably, The figures show axial slices of the multi-organ segmentation results. We compare the results with other models to show that our model outperforms others. When it comes to producing smooth boundaries, 3D-UNet shows its strength, but the prediction itself lacks accuracy. SwinUNETR seems more sensitive to intensity shifts within the image looking at

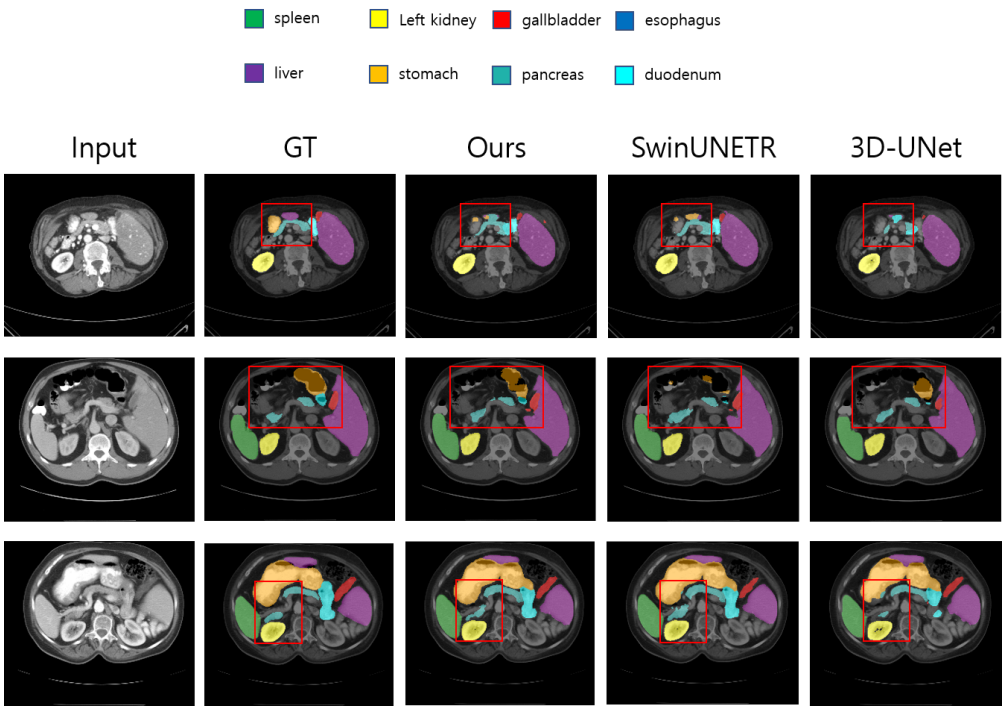


Figure 4.1: The axial slices of segmentation results

the prediction of the stomach of the second row input image. Our model is accurate compared to others and also have smoother boundaries than the other model with a ViT backbone.

4.4 Ablation Study

In this section, we investigate the effectiveness of the deformable attention blocks and ablate the key components in our proposed method.

Deformable attention blocks at different stages In order to verify the effectiveness of the proposed design, we applied deformable attention blocks instead of the shifted-window attention blocks at different stages. As shown in table 4.3, replacing the shifted-window blocks only at stage 3 and 4 leads to a

Stages w/ Deformable Attention Block				Avg. DSC
Stage 1	Stage 2	Stage 3	Stage 4	
O	O	O	O	0.784
	O	O	O	0.809
		O	O	**0.813**
O	O	O		0.791
O	O			0.793
SWIN-UNETR				0.796

Table 4.3: Ablation study on applying deformable attention blocks at different stages

performance gain of 0.3. Replacing shifted-window blocks in the earlier stages decreases the performance of the model.

Ablation on offset scales We conduct an experiment of using different maximum offset range scale factor to explore the robustness of our deformable attention block to this hyper-parameter. A wide range of values are applied as the offset range scale factor, ranging from 4 to 12 as 12 is the largest reasonable offset given in the size of the feature map($12 \times 12 \times 12$ at stage 4). Offset kernel sizes were fixed to [8, 8, 8] and [6, 6, 6] in stage 3 and 4 respectively. The results are shown in table 4.4.

Parameters	Avg. DSC
stage 3: Offset scale: [4, 4, 4] Offset kernel size: [8, 8, 8] stage 4: Offset scale: [4, 4, 4] Offset kernel size: [6, 6, 6]	**0.813**
stage 3: Offset scale: [8, 8, 8] Offset kernel size: [8, 8, 8] stage 4: Offset scale: [8, 8, 8] Offset kernel size: [6, 6, 6]	0.805
stage 3: Offset scale: [12, 12, 12] Offset kernel size: [8, 8, 8] stage 4: Offset scale: [12, 12, 12] Offset kernel size: [6, 6, 6]	0.761

Table 4.4: Ablation study on different offset range factor

Chapter 5

Discussion

Recent development in deep learning methods for multi-organ segmentation have explored various architectures, such as encoder-decoder based CNN networks [4, 14, 15] or ViT based networks [18, 21, 20, 19], for encoding high-level features using limited training data. However, CNN networks lack the ability to model long-ranged dependencies and transformers have quadratic computational complexity relative to the image size due to self-attention. Our proposed method is effective for reducing the computational complexity to be linear relative to the image size and also improve feature learning through the deformable attention module. A recent study [20] uses both CNNs and transformers to complement each other, but because of the computational burden, the input volume had to be sliced into 2D planes and later restored the volume through extra post-processing. Furthermore, in such a data dependently driven task and considering that certain organs have vague boundaries, applying our deformable attention module achieves more accurate segmentation results compared to previous works. Although self-attention is applicable in doing so, CNN

counterparts are better in learning local features and costs too much computational resources. There is still room for more extensive research in effectively and efficiently encoding local and global contexts as there are more studies in different partitioning schemes.

Chapter 6

Conclusion

In this work, we propose a new model for a multi-organ segmentation task. Our method learns more informative parts rather than focusing on the relation between patches and the entire image, which means that it preserves computational efficiency and encodes both local and global context for achieving precise segmentation results. Our method could focus more on the informative parts of the image in local windows by employing the deformable attention module, which is only a small addition to the computational complexity. The experimental results demonstrated that our method is either superior or similar to other methods based on CNNs or ViTs as backbones and also proves the effectiveness of our proposed method through ablation studies. Both quantitative and qualitative results also show that it is more effective in segmentation of smaller organs.

The limitations that we face in this study is the computational burden. In order to fit the process in our given environment, downsampling reference points was inevitable for our proposed method to work, which could severely

affect the results of multi-head attention after feature sampling based on the offsets. Further research can focus more on the coverage of features for global feature representation while maintaining the downsampling of reference points or developing new partition schemes that could effectively capture both global and local feature representations.

Bibliography

[1] G.-P. Glombitza, H. Evers, S. Haßfeld, U. Engelmann, and H.-P. Meinzer, "Virtual surgery in a (tele-)radiology framework," *IEEE Transactions on Information Technology in Biomedicine*, vol. 3, pp. 186–196, 1999.

[2] R. D. Howe and Y. Matsuoka, "Robotics for surgery," *Annual Review of Biomedical Engineering*, vol. 1, no. 1, pp. 211–240, 1999. PMID: 11701488.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems* (F. Pereira, C. Burges, L. Bottou, and K. Weinberger, eds.), vol. 25, Curran Associates, Inc., 2012.

[4] Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: Learning dense volumetric segmentation from sparse annotation," 2016.

[5] P. Ramachandran, N. Parmar, A. Vaswani, I. Bello, A. Levskaya, and J. Shlens, "Stand-alone self-attention in vision models," 2019.

[6] A. Parvaiz, M. A. Khalid, R. Zafar, H. Ameer, M. Ali, and M. M. Fraz, "Vision transformers in medical computer vision – a contemplative retrospection," 2022.

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2018.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017.

[9] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," 2021.

[10] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," 2017.

[11] Z. Xia, X. Pan, S. Song, L. E. Li, and G. Huang, "Vision transformer with deformable attention," 2022.

[12] R. Adams and L. Bischof, "Seeded region growing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, no. 6, pp. 641–647, 1994.

[13] Y. Boykov and M. Jolly, "Interactive graph cuts for optimal boundary and region segmentation of objects in n-d images," in *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, vol. 1, pp. 105–112 vol.1, 2001.

[14] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," 2016.

[15] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert, "Attention u-net: Learning where to look for the pancreas," 2018.

[16] M. Chung, J. Lee, S. Park, C. E. Lee, J. Lee, and Y.-G. Shin, "Liver segmentation in abdominal CT images via auto-context neural network and self-supervised contour attention," *Artificial Intelligence in Medicine*, vol. 113, p. 102023, mar 2021.

[17] S. Park and M. Chung, "Cardiac segmentation on ct images through shape-aware contour attentions," *Computers in Biology and Medicine*, vol. 147, p. 105782, 2022.

[18] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2020.

[19] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-unet: Unet-like pure transformer for medical image segmentation," 2021.

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

[21] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. Roth, and D. Xu, "Unetr: Transformers for 3d medical image segmentation," 2021.

[22] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. J. Cardoso, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pp. 240–248, Springer International Publishing, 2017.

[23] M. Yi-de, L. Qing, and Q. Zhi-bai, "Automated image segmentation using improved pcnn model based on cross-entropy," in *Proceedings of 2004 International Symposium on Intelligent Multimedia, Video and Speech Processing, 2004.*, pp. 743–746, 2004.

[24] E. Gibson, F. Giganti, Y. Hu, E. Bonmati, S. Bandula, K. Gurusamy, B. Davidson, S. P. Pereira, M. J. Clarkson, and D. C. Barratt, "Multi-organ Abdominal CT Reference Standard Segmentations," Feb. 2018. This data set was developed as part of independent research supported by Cancer Research UK (Multidisciplinary C28070/A19985) and the National Institute for Health Research UCL/UCL Hospitals Biomedical Research Centre.

[25] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2017.

[26] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. Roth, and D. Xu, "Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images," 2022.

초록

다중 장기 분할은 컴퓨터 보조 진단의 임상 응용에 중요한 작업이다. 딥러닝, 특히 합성곱 신경망의 최근 발전은 의료 영상에서 다중 장기의 동시 분할에 대한 유망한 결과를 보여주었다. 그러나 합성곱 신경망으로 이루어진 백본을 가진 대부분의 접근 방식은 합성곱 신경망의 제한된 모양으로 인해 전역 특정 표현이 약해질 우려가 있다. 다중 장기 분할을 위한 여러 아키텍쳐들이 탐구됨에 따라, 비전 트랜스포머(ViT)는 이미지 분류 작업에서 합성곱 신경망의 성능을 능가하는 상당한 개선을 보여주었다. 본 논문에서는 비정형 합성곱 신경망 네트워크에서 영감을 받아 다중 장기 분할의 성능을 높이기 위해 각 패치를 전체 이미지와 비교하는데 초점을 맞추기보다는 이미지의 중요 영역에 어텐션을 집중하기위해 오프셋과 척도를 학습하는 비정형 어텐션 네트워크를 소개한다. 전체 아키텍처는 합성곱 신경망과 비전 트랜스포머를 모두 효과적으로 활용하여 정확도를 높일 뿐만 아니라 셀프 어텐션 메커니즘의 계산 복잡성을 감소시킨다. 모델의 학습 및 검증을 위해 30개의 CT 복부 이미지를, 실제 테스트를 위해 17개의 CT 복부 이미지를 포함하는 BTCV(Beyond the Cranium Vault) 데이터 세트를 사용하였다. 실험 결과는 제안된 네트워크가 다이스 유사도 계수(DSC) 점수 측면에서 췌장 분할에서 3.7%, 십이지장 분할에서 3.6% 증가를 달성하여 이전 방법에 비해 더 정확한 결과를 생성한다는 것을 보여준다.

주요어: 심층 신경망, 비정형 어텐션, 다중 장기 분할, 의료 영상 분할

학번: 2020-25153

Acknowledgements

학부를 졸업한 후 컴퓨터공학 분야에서 전문성을 갖출 수 있도록 공부를 하겠다는 목표를 갖고 진학한 대학원이었습니다. 정말 운이 좋게 지금의 연구실에서 2년 조금 넘는 시간 동안 많은 것들을 경험하고 배울 수 있었고, 그 과정에서 많은 분들의 지도와 도움이 있었기에 이렇게 무사히 졸업할 수 있었던 것 같습니다. 이에 이 자리를 빌려 감사의 말씀을 전하고 싶습니다.

먼저 저의 지도교수님 신영길 교수님께 깊은 감사의 인사를 올립니다. 석사의 길을 걷도록, 뛰어난 역량의 사람들과 함께 연구할 수 있도록 기회를 주셔서 감사합니다. 항상 좋은 연구실 환경에서 연구와 개발을 할 수 있도록 신경 써주시고, 언제나 아낌없는 조언과 지도를 해주신 덕분에 연구실에 있는 동안 많은 것을 배우고 성장할 수 있었습니다. 교수님께서 해주신 조언들 잊지 않고 가슴에 새기도록 하겠습니다.

석사 기간 동안 연구실 생활을 함께한 연구실 선후배님들께 감사의 인사 드립니다. 제가 학부를 미국에서 나와서 많이 서투르고 부족했지만 제가 적응할 수 있도록 도와주시고 올바른 연구자의 표본으로 모범이 되어주신 민영이형께 감사하다는 말씀을 전하고 싶습니다. 프로젝트를 같이 진행하면서 부족한 저에게 많은 가르침을 주시고 항상 챙겨주셨던 민경누나께도 특별히 감사드립니다. 제가 잘 적응할 수 있도록 연구실의 분위기를 이끌어주셨던 지완이형, 항상 적극적으로 조언해주시고 큰 기둥처럼 든든했던 민창이형, 많은 시간 같이 못 보냈지만 멋지게

묵묵히 연구를 하시는 승환이형, 경휘누나께도 감사드립니다. 지금은 안 계시지만 처음에 적응하는데 큰 도움을 주셨던 지강이형, 그리고 지금은 다른 분야에서 꿈을 이루고 계신 민규형께도 감사드립니다. 연구실에서의 첫 프로젝트를 함께한 강용이형도 많은 것을 가르쳐주셔서 감사합니다. 처음 연구실에 들어왔을 때 제가 귀찮게 계속 질문해도 항상 친절하게 답해주고 제가 적응할 수 있도록 크게 도와준 옆자리와 뒷자리에 있던 상욱이와 주상이형께 감사드립니다. 항상 열정적으로 공부하고 뭐든지 잘해내 큰 자극이 되었던 채은이와 혜림이, 그리고 든든하게 도와주었던 호연이형에게도 감사드립니다.

항상 변함없는 믿음과 응원을 보내주고 사랑으로 키워주신 부모님께 감사 인사를 드립니다. 제가 많이 툴툴거리지만 항상 옳은 길로 인도해주시고 쓴소리가 필요할 때 해주셨습니다. 이제 곧 은퇴하시는 두분 모두 은퇴 후에 새로운 인생의 막을 의미있게 채워나가시길 바라며 하시는 일들 모두 잘되었으면 좋겠습니다. 저 멀리 미국에서 항상 응원과 고민을 같이 해주는 누나도 감사하다는 말 전합니다. 마지막으로 연구실 밖에서 학교 생활이 즐거울 수 있도록 해준 친구들과 임팩트원들 모두 너무 감사합니다.