

AFTer-UNet: Axial Fusion Transformer UNet for Medical Image Segmentation

Xiangyi Yan Hao Tang Shanlin Sun Haoyu Ma Deying Kong Xiaohui Xie
University of California, Irvine

{xiangyy4, htang6, shanlins, haoyum3, deyingk, xhx}@uci.edu

Abstract

Recent advances in transformer-based models have drawn attention to exploring these techniques in medical image segmentation, especially in conjunction with the U-Net model (or its variants), which has shown great success in medical image segmentation, under both 2D and 3D settings. Current 2D based methods either directly replace convolutional layers with pure transformers or consider a transformer as an additional intermediate encoder between the encoder and decoder of U-Net. However, these approaches only consider the attention encoding within one single slice and do not utilize the axial-axis information naturally provided by a 3D volume. In the 3D setting, convolution on volumetric data and transformers both consume large GPU memory. One has to either downsample the image or use cropped local patches to reduce GPU memory usage, which limits its performance. In this paper, we propose Axial Fusion Transformer UNet (AFTer-UNet), which takes both advantages of convolutional layers' capability of extracting detailed features and transformers' strength on long sequence modeling. It considers both intra-slice and inter-slice long-range cues to guide the segmentation. Meanwhile, it has fewer parameters and takes less GPU memory to train than the previous transformer-based models. Extensive experiments on three multi-organ segmentation datasets demonstrate that our method outperforms current state-of-the-art methods.

1. Introduction

Medical image segmentation is an essential procedure in many modern clinical workflows. It can be used in many applications, including diagnostic interventions, treatment planning and treatment delivery [15, 39]. These image analyses are usually carried out by experience doctors. However, it is labor-intensive and time-consuming, since a 3D CT volume can contain up to hundreds of 2D slices. Therefore, developing robust and accurate image segmentation tools is a fundamental need in medical image analysis [33, 34].

Traditional medical image segmentation methods are

mostly atlas-based. These methods usually rely on pre-computed templates, so they may not adequately account for the anatomical variance due to variations in organ shapes, removal of tissues, growth of tumor and differences in image acquisition. With the rise of deep learning, convolutional neural networks (CNNs) have been widely used in different domains of computer vision because of its extraordinary capability of extracting image features, such as object detection [30], semantic segmentation [24] and pose estimation [44, 43, 26], etc. U-Net [31] is the first to use CNNs in the field of medical image segmentation. Now U-Net and its variants [10, 52, 18] have achieved great success on this task.

Although CNNs are able to extract rich features, CNN-based approaches are not adequately equipped to encode long range interaction information [42], whether within one single slice (intra-slice) or among the neighboring slices (inter-slice). In the field of medical image segmentation, it is useful to capture this information, since the texture, shape and size of many organs vary greatly across patients and it often requires long-range contextual information to reliably segment these organs.

In the field of natural language processing (NLP), transformer-based methods [41] have achieved the state-of-the-art-performance in many tasks. Inspired by this design, researchers naturally think of leveraging Transformers' ability of modeling long range relationships to improve pure CNN-based models in natural images. However, less attention has been paid to use transformer-based models in medical image segmentation.

Recently, transformer-based models have been proposed in medical image segmentation in both 2D and 3D settings, with pros and cons associated with each as follows. In the 2D setting, TransUNet [7] is the first to investigate the usage of Transformers for medical image segmentation to model long-range dependencies within a single 2D image. However, it does not consider the long-range dependencies in the 3D data, *i.e.*, along the axial-axis, which is naturally provided in the 3D medical image data [36]. In the 3D setting, CoTr [45] is the first to explore Transformers to model long-range relationships in the volumetric data. However,

because transformer modules and volumetric data both consume a lot of GPU memory, they need to compromise both in order to fit their model into easily accessible commodity GPUs. To address this, they cut the 3D volumetric data into local patches and process them one at a time, which results in loss of information from other patches. Moreover, they limit the pairwise attention to only a few voxels in the 3D data, which may be oversimplified and limit the ability of Transformers in modeling long-range relationships.

To better utilize Transformer to explore the long-range relationships in the 3D medical image data, in this paper, we propose **Axial Fusion Transformer UNet (AFTER-UNet)**, an end-to-end medical image segmentation framework. Our motivation is to leverage both intra-slice and inter-slice contextual information to guide the final segmentation step. AFTER-UNet follows the U-shape structure of U-Net, which contains a 2D CNN encoder and a 2D CNN decoder. In between, we propose axial fusion transformer encoder to fuse contextual information in the neighboring slices. The axial fusion transformer encoder reduces the computational complexity by first separately calculating the attention along the axial axis and the attention within one single slice, and then fusing them together to produce the final segmentation map.

Our main contributions are listed as follows:

- We propose an end-to-end framework, Axial Fusion Transformer UNet, to deal with 3D medical image segmentation tasks by fusing intra-slice and inter-slice information.
- We introduce axial fusion mechanism, which reduces the computational complexity of calculating self-attention in 3D space.
- We conduct extensive experiments on three multi-organ segmentation benchmarks, and demonstrate superior performance of AFTER-UNet compared to current transformer based models.

2. Related work

2.1. CNN-based segmentation networks

Early medical image segmentation methods are mainly contour-based and traditional machine learning-based algorithms. With the development of deep CNN, U-Net is proposed in [31] for medical image segmentation. Due to the simplicity and superior performance of the U-shaped structure, various UNet-like methods are constantly emerging, such as Res-UNet [10], Dense-UNet [5], U-Net++ [52] and UNet3+ [18]. And it is also introduced into the field of 3D medical image segmentation, such as 3D U-Net [53] and V-Net [27]. It is also extended to other medical image analysis tasks, such as computer-aided diagnosis [35, 40, 38, 23, 11], image denoising [48, 25], image registration [2, 17], etc. At present, CNN-based methods

have achieved tremendous success in the field of medical image segmentation due to its powerful representation ability [19, 37, 14, 28, 13, 46, 47, 49].

2.2. Visual transformers

Transformer was first proposed for the machine translation task in [41]. In the domain of natural language processing, the Transformer-based methods have achieved the state-of-the-art performance in various tasks [9, 50]. Motivated by the success of [9], researchers introduced vision transformers (ViT) in [12] for image classification tasks. Besides, [3] extended ViT to the field of video classification, which largely inspired our work. For object detection, [6] predicts the final set of detections by combining a common CNN with a transformer architecture. For semantic segmentation, [51] exploited the transformer framework to implement the feature representation encoder by sequentializing images without using the traditional FCN [24] design.

2.3. Transformers for medical image segmentation

Recently, researchers have tried to apply transformer modules to improve the performance of current approaches. TransUNet [7] is the first paper to investigate the usage of Transformers for medical image segmentation problems. In this paper, the encoder and decoder of U-Net is connected by several Transformer layers. TransUNet leverages both CNN's capability of extracting low level features and Transformer's advantage of making high level sequence-to-sequence prediction. Swin-UNet [4] explores the application potential of pure transformer in medical image segmentation. However, both TransUNet and Swin-UNet only consider a single slice as input, so that the information along the axial-axis, which is intrinsically provided by a 3D volume, is not utilized. On the other hand, researchers have been trying applying transformers in a 3D way. However, computing self-attention directly on 3D space is not feasible due to the expensive computation. To resolve this issue, CoTr [45] introduces deformable self-attention mechanism, which indeed reduces the computational complexity. However, the design of CoTr brings two issues. First, it requires 3D patches as inputs, which means a lot of information are lost due to the split of patches. Actually, this is a common issue for 3D medical image segmentation models. Second, CoTr computes self-attention over only K locations. In their experiments, a larger K leads to a higher Dice score. However, K is only set to 4 as the maximum value. This is because the transformer module has to compromise more memory space for the expensive 3D convolutions. Therefore, the over-simplified way of computing self-attention in 3D space, may cost the loss contextual information. In our following experiments, CoTr shows marginal accuracy improvement compared to previous methods.

3. Method

Figure 1 shows the details of AFter-UNet. We follow the classic U-Net design, which includes a 2D CNN encoder for extracting fine-level image features and a 2D CNN decoder for achieving pixel-level segmentation. To better encode high-level semantic information, not only within a single slice but also among neighboring slices, we propose the axial fusion transformer in between. We now elaborate the details of each module in the following subsections.

3.1. CNN encoder

3.1.1 Input formulation

Given an input 3D CT scan $\mathbf{s} \in \mathbb{R}^{C \times H \times W \times D}$, we have a series of 2D slices along the axial axis, with height of H , width of W and channel of $C = 1$. The scan can be represented as $\mathbf{s} = \{\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_D\}$, where $\mathbf{s}_d \in \mathbb{R}^{C \times H \times W}$. Firstly, for each 2D slice \mathbf{s}_i , we sample its N_A neighboring slices along the axial axis by frequency N_f and get $\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_D\}$, where $\mathbf{x}_d \in \mathbb{R}^{C \times H \times W \times N_A}$. For the each sampled neighboring slice group \mathbf{x}_d , we have $\mathbf{x}_d = \{\mathbf{s}_{a_0}, \mathbf{s}_{a_1}, \dots, \mathbf{s}_{a_{N_A}}\}$, where $a_n = d - N_f \times (\frac{N_A}{2} - n)$ and $n \in \{0, 1, \dots, N_A\}$.

3.1.2 Architecture

The CNN encoder \mathcal{E}^{CNN} mainly follows the design of U-Net, which includes B blocks, connected by MaxPooling layers with both kernel size and stride of 2. Each block contains two Conv2d-ReLU pairs. Additionally, we also add instance normalization layers in each pair between the 2d convolution layer and the ReLU layer. Given an input neighboring slice group $\mathbf{x}_d \in \mathbb{R}^{C \times H \times W \times N_A}$, each CNN encoder block \mathcal{E}_b^{CNN} provides a corresponding feature map group $\mathbf{g}_{d,b} = \{\mathbf{g}_{d,b}^0, \mathbf{g}_{d,b}^1, \dots, \mathbf{g}_{d,b}^{N_A}\}$, where $\mathbf{g}_{d,b}^n$ denotes a feature map at level b for slice n and $\mathbf{g}_{d,b}^n \in \mathbb{R}^{C_b \times H_b \times W_b}$. Here we have $b \in \{1, 2, \dots, B\}$ indicating the feature level, $H_b = \frac{H}{2^b}$, $W_b = \frac{W}{2^b}$ and C_b denoting the height, width and number of channels at level b .

However, here we only take the final feature map group $\mathbf{g}_{d,B} = \mathcal{E}^{CNN}(\mathbf{x}_d)$ as input to the axial fusion transformer. We denote it as \mathbf{g} , where $\mathbf{g} \in \mathbb{R}^{C_L \times H_L \times W_L \times N_A}$, for simplicity. We choose this design for the following two reasons: First, taking the higher level feature map group means leveraging higher level semantic information, which is the motivation of applying transformers. Second, the GPU memory limits the size of the feature map group.

3.2. Axial fusion transformer encoder

After extracting fine level features by the CNN encoder, we now introduce the Axial Fusion Transformer encoder to model high level semantic information not only within a single slice but also among neighboring slices along the axial axis.

3.2.1 Feature maps as input embeddings

In [16], not matter the input is a 2D slice/image or a 3D volume, it needs to be divided into small patches and then linearly mapped to vectors of a certain length. This is because transformers can't handle inputs with large size due to the memory limit. In our design, with the help of the above CNN encoder, we now have each feature map as an input with much smaller height H_L and width W_L , so that each feature map \mathbf{g}_n can be directly fed into the axial fusion transformer. Meanwhile, our approach provides more comprehensive information extracted from the whole image than from a single patch.

We then directly have $\mathbf{z}_{((h,w),n)}^{(0)} \in \mathbb{R}^{C_L}$ without the linear projection step in the original ViT [12] setup:

$$\mathbf{z}_{((h,w),n)}^{(0)} = \mathbf{g}_{((h,w),n)} + \mathbf{e}_{((h,w),n)}^{pos} \quad (1)$$

, where $\mathbf{e}_{((h,w),n)}^{pos} \in \mathbb{R}^{C_L}$ represents a learnable positional embedding to encode the vector location: (1) at (h, w) within a single feature map \mathbf{g}_n and (2) at n among feature maps in group \mathbf{g} . The resulting sequence $\mathbf{z}_{((h,w),n)}^{(0)}$ for $(h, w) = (1, 1), \dots, (H_L, W_L)$ and $n = 0, 1, \dots, N_A$ represents the input to the Transformer, and plays a role similar to the sequences of embedded words that are fed to text Transformers in NLP. Note that in our code implementation, the dimensions of height and width are flatten so the vector can also be represented as $\mathbf{z}_{(p,n)}^{(0)}$, where $p = W_L \cdot (h - 1) + w$, but for illustration purpose, we keep the notion of (h, w) here.

3.2.2 Query-Key-Value matrices and self-attention

The axial fusion transformer consists of L blocks. For each block l , we compute each location's query/key/value vector from the representation $\mathbf{z}_{((h,w),n)}^{(l-1)}$, which is encoded by the preceding block:

$$\mathbf{q}_{((h,w),n)}^{(l,a)} = W_Q^{(l,a)} \text{LN} \left(\mathbf{z}_{((h,w),n)}^{(l-1)} \right) \in \mathbb{R}^{C_h} \quad (2)$$

$$\mathbf{k}_{((h,w),n)}^{(l,a)} = W_K^{(l,a)} \text{LN} \left(\mathbf{z}_{((h,w),n)}^{(l-1)} \right) \in \mathbb{R}^{C_h} \quad (3)$$

$$\mathbf{v}_{((h,w),n)}^{(l,a)} = W_V^{(l,a)} \text{LN} \left(\mathbf{z}_{((h,w),n)}^{(l-1)} \right) \in \mathbb{R}^{C_h} \quad (4)$$

, where $\text{LN}()$ denotes LayerNorm [1], $a \in \{1, 2, \dots, \mathcal{A}\}$ represents the index over multiple attention heads and \mathcal{A} denotes the total number of attention heads. Therefore, we have the dimensionality for each attention head as $C_h = C_L / \mathcal{A}$.

Self-attention weights are computed via dot-product. The self-attention weights $\alpha_{((h,w),n)}^{(l,a)} \in \mathbb{R}^{(H_L \cdot W_L) \cdot N_A}$ for query at $((h, w), n)$ are:

$$\alpha_{((h,w),n)}^{(l,a)} = \text{SoftMax} \left(\frac{\mathbf{q}_{((h,w),n)}^{(l,a)} \cdot \mathbf{k}_{((h,w)',n')}^{(l,a) \top}}{\sqrt{C_h}} \right) \quad (5)$$

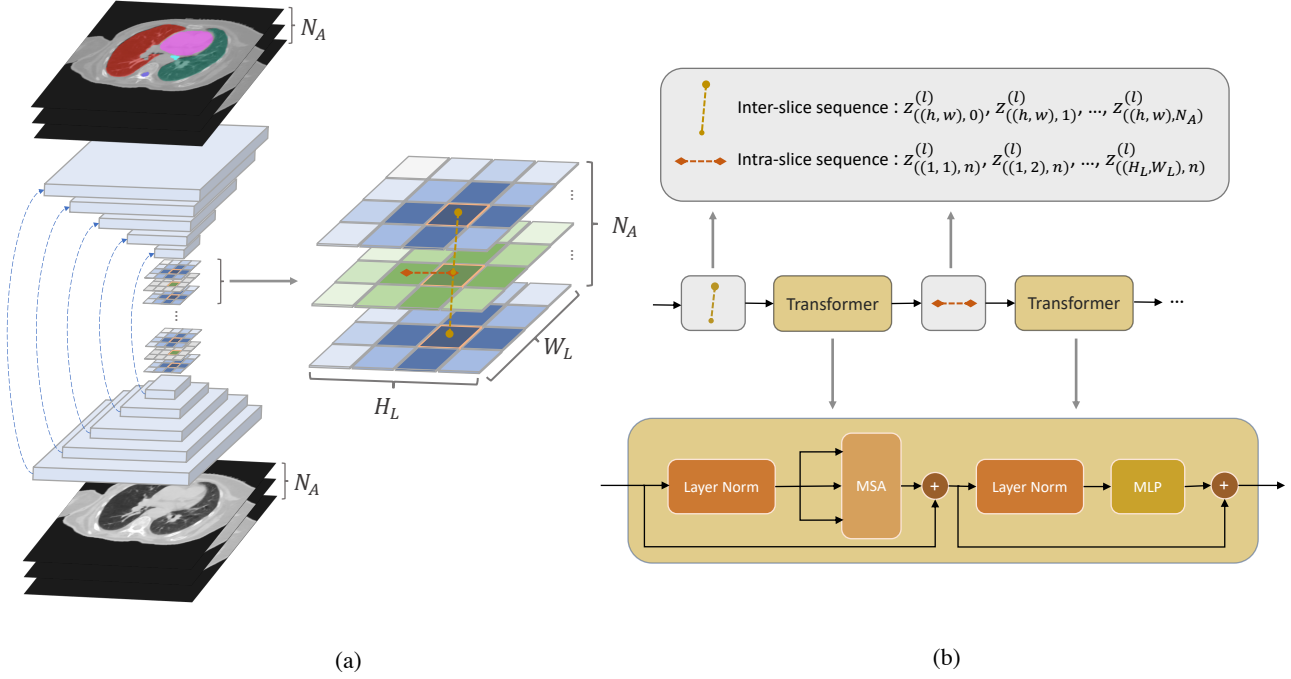


Figure 1. Overview of Axial Fusion Transformer UNet. (a) The architecture of AFTer-UNet. We first encode the neighboring slice group \mathbf{x}_d by the CNN encoder and get a corresponding feature map group \mathbf{g} . Then we apply axial fusion transformer to \mathbf{g} . Last, the feature group fused with both intra-slice and inter-slice cues are fed to the CNN decoder for segmentation. (b) The illustration of axial fusion mechanism. AFTer-UNet separately fuse the inter-slice and intra-slice information by Eq.6 and Eq.7 respectively.

, where $(h, w)' \in \{(1, 1), \dots, (H_L, W_L)\}$ and $n' \in \{0, 1, \dots, N_A\}$. Note that when attention is computed only within a single feature map or only along the axial axis, the computation is significantly reduced. In the case of computing attention within a single feature map, only $H_L \cdot W_L$ query-key comparisons are made, using exclusively keys from the same feature map as the query:

$$\alpha_{((h,w),n)}^{(l,a)\text{intra}} = \text{SoftMax}\left(\frac{\mathbf{q}_{((h,w),n)}^{(l,a)\top}}{\sqrt{C_h}} \cdot \mathbf{k}_{((h,w)',n)}^{(l,a)}\right) \quad (6)$$

, where $(h, w)' \in \{(1, 1), \dots, (H_L, W_L)\}$.

To get the encoding $\mathbf{z}_{((h,w),n)}^{(l)}$ at block l , we firstly compute the weighted sum of value vectors, using self-attention coefficients from each attention head:

$$\mathbf{u}_{((h,w),n)}^{(l,a)} = \sum_{(h,w)'=(1,1)}^{(H_L,W_L)} \sum_{n'=1}^{N_A} \alpha_{((h,w),n),((h,w)',n')}^{(l,a)} \mathbf{v}_{((h,w)',n')}^{(l,a)}. \quad (7)$$

These vectors from all heads are then concatenated, linearly projected by an fully connected layer (FC) and passed through an multi-layer perceptron (MLP) with layer norm (LN). Residual connections are added after each operation:

$$\mathbf{z}'_{((h,w),n)}^{(l)} = \text{FC} \left(\begin{array}{c} \mathbf{u}_{((h,w),n)}^{(l,1)} \\ \vdots \\ \mathbf{u}_{((h,w),n)}^{(l,A)} \end{array} \right) + \mathbf{z}_{((h,w),n)}^{(l-1)} \quad (8)$$

$$\mathbf{z}_{((h,w),n)}^{(l)} = \text{MLP} \left(\text{LN} \left(\mathbf{z}'_{((h,w),n)}^{(l)} \right) \right) + \mathbf{z}'_{((h,w),n)}^{(l)} \quad (9)$$

3.2.3 Fusing axial information

Due to the limit of memory, computing self-attention over a 3D space by Eq.5 is not feasible. Replacing it with 2D attention applied only on one single slice, *i.e.*, Eq.6 can certainly reduce the computational cost. However, such a model ignores to capture information among neighboring slices, which is naturally provided by a 3D volume. As shown in our experiments, considering less neighboring slices can provide poorer results.

We propose axial fusion mechanism for computing attention along the axial axis, where the attention along the axial axis and the attention within a single slice are separately applied one after the other. By fusing the axial information this way, we firstly compute attention along the axial with all the channels at the same position at (h, w) :

$$\alpha_{((h,w),n)}^{(l,a)\text{inter}} = \text{SoftMax}\left(\frac{\mathbf{q}_{((h,w),n)}^{(l,a)\top}}{\sqrt{C_h}} \cdot \mathbf{k}_{((h,w),n')}^{(l,a)}\right) \quad (10)$$

Methods	DSC	Eso	Trachea	Spinal Cord	Lung(L)	Lung(R)	Heart
U-Net[31]	91.18	78.85	90.72	89.37	97.31	96.37	94.46
nnUNet-2D[19]	89.74	78.82	88.32	86.61	96.03	96.65	92.01
nnUNet-3D[19]	91.63	81.18	89.32	91.21	97.68	97.74	92.66
Attention U-Net[29]	90.19	76.35	88.14	89.43	97.65	97.87	91.68
TransUNet[7]	91.38	78.27	91.45	88.36	97.63	97.84	94.74
Swin-Unet[4]	91.26	78.98	91.20	88.64	97.64	97.79	93.30
CoTr[45]	91.39	79.06	91.55	88.67	97.47	97.65	93.92
AFTer-UNet	92.32	81.47	91.76	90.12	97.80	97.90	94.86

Table 1. Dice scores of different methods on in-house thorax-85 dataset.

, where $n' \in \{1, \dots, N_A\}$. The encoding $\mathbf{z}'_{((h,w),n)}^{(l)\text{inter}}$ resulting from the application of Eq.8 using axial attention is then fed back for single slice attention computation instead of directly being passed to the MLP. In other words, new key/query/value vectors are obtained from $\mathbf{z}'_{((h,w),n)}^{(l)\text{inter}}$ and the single slice attention is then computed using Eq.6. Finally, the resulting vector $\mathbf{z}'_{((h,w),n)}^{(l)\text{intra}}$ is passed to the MLP of Eq.9 to compute the final encoding $\mathbf{z}_{((h,w),n)}^{(l)}$ at position $((H, W), n)$ by block l . The final fused encoding for the feature map group \mathbf{g} is $\mathbf{z}^{(L)} \in \mathbb{R}^{C \times H_L \times W_L \times N_A}$.

We learn distinct query/key/value matrices $\{W_{Q^{\text{slice}}}^{(l,a)}, W_{K^{\text{slice}}}^{(l,a)}, W_{V^{\text{slice}}}^{(l,a)}\}$ and $\{W_{Q^{\text{axial}}}^{(l,a)}, W_{K^{\text{axial}}}^{(l,a)}, W_{V^{\text{axial}}}^{(l,a)}\}$ over dimensions within one single slice and among slices along the axial axis. Note that compared to the $(H_L \cdot W_L) \cdot N_A$ comparisons each vector needed by the self-attention model of Eq.5, our approach performs only $(H_L \cdot W_L) + N_A$ comparisons per vector.

3.3. CNN decoder

The CNN decoder \mathcal{D}^{CNN} of AFTer-UNet follows the design of U-Net as well, which is mostly symmetric to the CNN encoder \mathcal{E}^{CNN} . It includes B Conv2d-ReLU blocks. Adjacent blocks are connected by Upsample layers with the scale factor of 2. Each block contains two Conv2d-ReLU pairs, with instance normalization layers between the 2d convolution layer and the ReLU layer.

The final fused feature map group $\mathbf{z}^{(L)}$, provided by the last block of axial fusion transformer encoder, is taken as input to the CNN decoder. It gets progressively upsampled to the input resolution by the Upsample layers and gets refined by the Conv2d-ReLU blocks. As applied in the U-Net paper, we also the skip connections between encoder and decoder to keep more low-level details for better segmentation.

Therefore, taking a series of sampled neighboring slice groups \mathbf{x} as inputs, we now have d segmentation map groups $\mathbf{y} = \{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_D\}$ as outputs, where $\mathbf{y}_d \in \mathbb{R}^{C_{cls} \times H \times W \times N_A}$ and C_{cls} denotes the number of organ classes. We only keep the middle segmentation map $\mathbf{y}_d^{\frac{N_A}{2}}$,

remove its neighbor for all d segmentation map groups and concatenate them together. At last, we have the final 3D prediction with respect to the 3D scan.

The loss function of our model is the sum of the dice loss and cross entropy loss.

4. Experiments

4.1. Setup

Dataset We conducted experiments using one abdomen CT dataset and two thorax CT datasets:

- BCV is in the MICCAI 2015 Multi-Atlas Abdomen Labeling Challenge [22]. It contains 30 3D abdominal CT scans from patients with various pathologies and has variations in intensity distributions between scans. Following [7, 4], we report the average DSC on 8 abdominal organs (aorta, gallbladder, spleen, left kidney, right kidney, liver, pancreas, spleen, stomach) with a random split of 18 training cases and 12 test cases.

- Thorax-85 is an in-house dataset from [8] that contains 85 3D thorax CT images. We report the average DSC on 6 thorax organs (eso, trachea, spinal cord, left lung, right lung, and heart) with a random split of 60 training cases and 25 test cases.

- SegTHOR is from the 2019 Challenge on Segmentation of THoracic Organs at Risk in CT Images [21]. It contains 40 3D thorax CT scans. We report the average DSC on 4 thorax organs (eso, trachea, aorta, and heart) with a random split of 30 training cases and 10 validation cases.

Evaluation metric We use the same evaluation metric Sørensen–Dice coefficient (DSC) as in previous work [7, 45]. DSC measures the overlap of the prediction mask \mathbf{m}_p and ground truth mask \mathbf{m}_g and is defined as:

$$\text{DSC}(\mathbf{m}_p, \mathbf{m}_g) = \frac{2|\mathbf{m}_p \cup \mathbf{m}_g|}{|\mathbf{m}_p| + |\mathbf{m}_g|} \quad (11)$$

Implementation details All images are resampled to have spacing of $2.5\text{mm} \times 1.0\text{mm} \times 1.0\text{mm}$, with respect to the depth, height, and width of the 3D volume. In the training stage, we apply elastic transform for alleviating overfitting. We use Adam[20] optimizer with momentum of 0.9 and weight decay of 10^{-4} to train AFTer-UNet for

Methods	DSC	Aorta	Gallbladder	Kidney(L)	Kidney(R)	Liver	Pancreas	Spleen	Stomach
U-Net[31]	74.68	87.74	63.66	80.60	78.19	93.74	56.90	85.87	74.16
Attention U-Net[29]	75.57	55.92	63.91	79.20	72.71	93.56	49.37	87.19	74.95
TransUNet[7]	77.48	87.23	63.13	81.87	77.02	94.08	55.86	85.08	75.62
Swin-Unet[4]	79.13	85.47	66.53	83.28	79.61	94.29	56.58	90.66	76.60
CoTr[45]	78.46	87.06	63.65	82.64	78.69	94.06	57.86	87.95	75.74
AFTer-UNet	81.02	90.91	64.81	87.90	85.30	92.20	63.54	90.99	72.48

Table 2. Dice scores of different methods on the Synapse multi-organ CT (BCV) dataset.

Methods	DSC	Eso	Trachea	Aorta	Heart
U-Net	89.97	80.07	91.23	94.73	93.83
Att U-Net	90.47	81.25	90.82	94.74	95.07
TransUNet	91.50	81.41	94.05	94.48	96.07
Swin-Unet	91.29	81.06	93.27	94.82	96.02
CoTr	91.41	81.53	94.03	94.06	96.01
AFTer-UNet	92.10	82.98	94.20	94.92	96.31

Table 3. Dice scores of different methods on SegTHOR thorax dataset.

550 epochs. The learning rate is set to 10^{-4} for the first 500 epochs and 10^{-5} for the last 50 epochs. In one epoch, for each 3D CT scan s , we only randomly select one slice group x_d , rather than all of them. We set the number of Conv2d-ReLU blocks $B = 5$, number of axial fusion transformer $L = 6$, number of attention heads $\mathcal{A} = 8$, number of neighboring slices $N_A = 8$ and sample frequency $N_f = 1$.

4.2. Results on Thorax-85

Table 1 shows the performance comparison of AFTer-UNet with previous work on Thorax-85. We ran the following representative algorithms: U-Net [31], Attention U-Net [29], nnU-Net [19], TransUNet [7], Swin-Unet [4], and CoTr [45]. U-Net is a well-established medical image segmentation baseline algorithm. Attention U-Net [29] is a multi-organ segmentation framework that uses gated attention to filter out irrelevant responses in the feature maps. nnU-Net [19] is a self-adaptive medical image semantic segmentation framework that wins the first in the Medical Segmentation Decathlon(MSD) challenge [32]. TransUNet [7] presents the first study which explores the potential of transformers in the context of 2D medical image segmentation. Swin-Unet [4] explores using pure transformer modules on 2D medical image segmentation tasks, without any convolutional layers. CoTr [45] firstly explores transformer modules for 3D medical image segmentation. The above-mentioned works cover a wide range of algorithms for multi-organ segmentation and should provide a comprehensive and fair comparison to our method on the in-house Thorax-85 dataset.

By comparing the results on left lung, right lung and heart, all models provide comparable results. This is because those organs are usually large and have regular shapes. However, for organs like esophagus and trachea, 3D models and transformer-based models have consistently higher DSC. These organs often have more anatomical variance, so the capability of long-range sequence modeling provides a holistic understanding of the context, which is beneficial. Both CoTr and AFTer-UNet consider using transformers to fuse 3D information. However, CoTr directly takes 3D patches as inputs, which may cause the loss of spatial information inter-patches. Besides, due to CoTr’s heavy 3D convolution operation, they limit the pairwise attention to only a few voxels, which may be oversimplified and limit the ability of Transformers in modeling long-range relationships. AFTer-UNet, however, applies 2D convolution to extract fine-level detail features and leave more memory space for the axial fusion transformer to extract richer inter-slice and intra-slice information. On our in-house thorax-85 dataset, the higher capability of long dependency modeling enables AFTer-UNet to outperform the previous state-of-the-art transformer-based method CoTr by 0.95%. Altogether, we demonstrated the effectiveness of the proposed method, which achieves an average DSC of 92.32% on six thorax organs.

4.3. Results on public datasets

We also conduct experiments on a public abdomen dataset, BCV, and a public thorax dataset, SegTHOR. Table 2 and 3 show the performance of AFTer-UNet and previous models. For large and normal-shaped organs, such as liver, spleen, stomach, and heart, all models are on par with each other. This is consistent with the conclusion we draw in section 4.2. However, for organs like aorta, left kidney, right kidney and pancreas, AFTer-UNet outperforms U-Net baseline by 3.17%, 7.30%, 7.11%, and 6.64% respectively and outperforms CoTr by 3.85%, 5.26%, 6.61% and 5.68% respectively. On average, AFTer-UNet has 4.34% improvement compared to U-Net and 2.56% improvement compared to CoTr. For elongated shaped organs like the esophagus in SegTHOR, AFTer-UNet provides 2.91% improvement compared to U-Net baseline and 1.45% improvement compared to CoTr. On average, our AFTer-UNet outper-

N_A	DSC	Eso	Trachea	Spinal Cord	Lung(L)	Lung(R)	Heart
1	91.44	78.31	90.65	90.14	97.48	97.69	94.35
2	91.66	78.54	91.35	90.27	97.59	97.60	94.59
4	91.98	79.74	91.42	90.71	97.59	97.77	94.66
8	92.32	81.47	91.76	90.12	97.80	97.90	94.86

Table 4. Ablation study on N_A , the number of neighboring axial slices.

L	DSC	Eso	Trachea	Spinal Cord	Lung(L)	Lung(R)	Heart
1	91.19	80.47	91.38	87.6	96.43	96.38	94.89
2	92.13	80.54	91.40	90.64	97.63	97.76	94.82
4	92.25	80.66	91.66	90.78	97.70	97.83	94.88
6	92.32	81.47	91.76	90.12	97.80	97.90	94.86

Table 5. Ablation Study on L , the number of transformer layers.

N_f	DSC	Eso	Trachea	Spinal Cord	Lung(L)	Lung(R)	Heart
1	92.32	81.47	91.76	90.12	97.80	97.90	94.86
2	91.92	79.77	91.06	90.49	97.69	97.65	94.87
4	92.05	79.72	91.69	90.30	97.81	97.88	94.89

Table 6. Ablation study on N_f , the sampling frequency on axial axis.

forms U-Net by 2.13% and CoTr by 0.69%.

4.4. Ablation study on Thorax-85

We further conduct extensive ablation studies on Thorax-85 to explore the influence of different hyperparameters:

The number of neighboring axial slices N_A . As discussed in previous sections, AFTer-UNet fuses inter-slice information by using the axial fusion mechanism. The number of neighboring axial slices is an essential factor of the mechanism. It is observed from Table 4 that the more neighboring slices we fuse, the higher the dice score is. Note that for elongated shaped organs such as esophagus and trachea, the dice scores increase more obviously than other large and normal shaped organs such as left lung, right lung and heart. It makes sense because those organs with large anatomical variances require more global information and increasing the number of neighboring slices can help to fulfill this requirement. In our AFTer-UNet model, we use $N_A = 8$ as the number of neighboring axial slices.

The number of axial fusion transformer layers L . Table 5 shows the results with $L = 1, 2, 4, 6$. It is observed that the average dice scores are improved when the number of axial fusion transformer layers goes up. Especially for elongated shaped organs such as esophagus and trachea, as L increases, dice scores on these two organs are improved more obviously than other large organs with normal shapes. This again shows the effectiveness of the axial fusion mechanism on compounding inter-slice cues.

The sampling frequency on axial axis N_f . We conduct experiments with various $N_f = 1, 2, 4$, and results

are shown in Table 6. It turns out that increasing the sampling frequency will hurt AFTer-UNet’s performance. In our design of axial fusion mechanism, lower N_f leads to **denser** inter-slice information. The results show that it’s more important to fuse the nearest neighboring information than slices far away. This might be one of the reasons why AFTer-UNet outperforms CoTr. The latter only considers several key points information, which are **sparsely** distributed in a 3D volume.

4.5. Qualitative results

Fig.2 shows the qualitative results of different approaches on Thorax-85 dataset. Thanks to the axial fusion mechanism, AFTer-UNet presents its effectiveness compared to other methods. We may focus on esophagus, the hardest to segment in Thorax-85, due to its large anatomy variance and elongated shape. For previous methods, segmentation maps might move largely between slices, which is unreasonable. However, AFTer-UNet (last column) provides consecutive and accurate predictions by considering the inter-slice and intra-slice context. Note that in this part, we didn’t visualize trachea since 1) trachea is relatively naive to segment so all model provides accurate results and 2) trachea lies in the very top region of thorax where few other organs can be shown at the same time.

4.6. Memory consumption and model parameters

We compare the GPU memory consumption of 1) directly computing self-attention on 3D and 2) computing by our proposed axial fusion mechanism. As shown in Fig.3, our proposed axial fusion mechanism leads to dra-

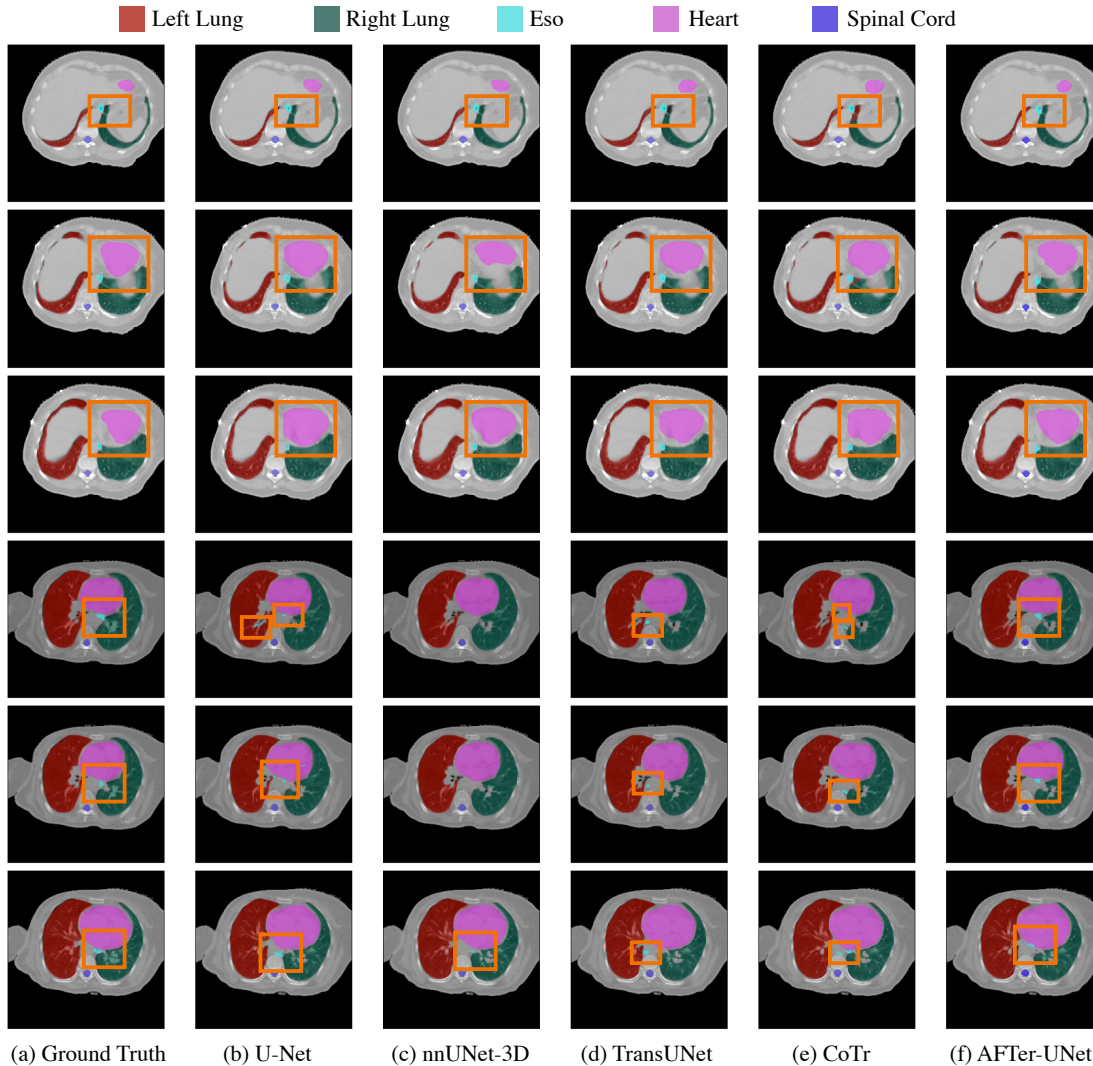


Figure 2. Qualitative results of different approaches on Thorax-85 dataset. (a) shows the ground truth of the CT slice. (b)-(e) show the results of previous methods. (f) shows the results of AFTer-UNet. The regions in orange rectangles indicate the effectiveness to our model.

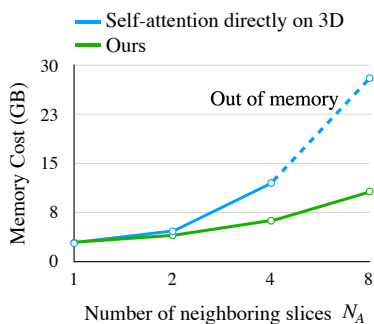


Figure 3. GPU memory consumption comparison between computing self-attention on 3D directly and computing by our proposed axial fusion mechanism.

matic computational savings. Therefore, our AFTer-UNet is

able to be trained on a single RTX-2080Ti GPU with 11GB memory. Besides, TransUNet has 43.5M parameters and CoTr has 41.9M parameters [45]. Meanwhile, AFTer-UNet has 41.5M parameters. This shows our method doesn't include more parameters to achieve its effectiveness than previous transformer based models.

5. Conclusion

In this work, we introduce AFTer-UNet, an end-to-end framework for medical image segmentation. The proposed framework use an axial fusion mechanism to fuse intra-slice and inter-slice contextual information and guide the final segmentation process. Experiments on three datasets demonstrate our model's effectiveness compared to previous work.

References

- [1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization, 2016.
- [2] Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca. Voxelmorph: a learning framework for deformable medical image registration. *IEEE transactions on medical imaging*, 38(8):1788–1800, 2019.
- [3] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proceedings of the International Conference on Machine Learning (ICML)*, July 2021.
- [4] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation, 2021.
- [5] Yue Cao, Shigang Liu, Yali Peng, and Jun Li. Denseunet: densely connected unet for electron microscopy image segmentation. *IET Image Processing*, 14(12):2682–2689, 2020.
- [6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers, 2020.
- [7] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L. Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.
- [8] Xuming Chen, Shanlin Sun, Narisu Bai, Kun Han, Qianqian Liu, Shengyu Yao, Hao Tang, Chupeng Zhang, Zhipeng Lu, Qian Huang, Guoqi Zhao, Yi Xu, Tingfeng Chen, Xiaohui Xie, and Yong Liu. A deep learning-based auto-segmentation system for organs-at-risk on whole-body computed tomography images for radiation therapy. *Radiotherapy and Oncology*, 160:175–184, July 2021.
- [9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.
- [10] Foivos I. Diakogiannis, François Waldner, Peter Caccetta, and Chen Wu. Resunet-a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 162:94–114, Apr 2020.
- [11] Jia Ding, Aoxue Li, Zhiqiang Hu, and Liwei Wang. Accurate pulmonary nodule detection in computed tomography images using deep convolutional neural networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 559–567. Springer, 2017.
- [12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.
- [13] Tianming Du, Yanci Zhang, Xiaotong Shi, and Shuang Chen. Multiple slice k-space deep learning for magnetic resonance imaging reconstruction. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, pages 1564–1567, 2020.
- [14] Yunhe Gao, Rui Huang, Ming Chen, Zhe Wang, Jincheng Deng, Yuanyuan Chen, Yiwei Yang, Jie Zhang, Chanjuan Tao, and Hongsheng Li. Focusnet: Imbalanced large and small organ segmentation with an end-to-end deep neural network for head and neck ct images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 829–838. Springer, 2019.
- [15] Eli Gibson, Francesco Giganti, Yipeng Hu, Ester Bonmati, Steve Bandula, Kurinchi Gurusamy, Brian Davidson, Stephen P Pereira, Matthew J Clarkson, and Dean C Barratt. Automatic multi-organ segmentation on abdominal ct with dense v-networks. *IEEE transactions on medical imaging*, 37(8):1822–1834, 2018.
- [16] Ali Hatamizadeh, Dong Yang, Holger Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation, 2021.
- [17] Mattias P Heinrich. Closing the gap between deep and conventional image registration using probabilistic dense displacement networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 50–58. Springer, 2019.
- [18] Huimin Huang, Lanfen Lin, Ruofeng Tong, Hongjie Hu, Qiaowei Zhang, Yutaro Iwamoto, Xianhua Han, Yen-Wei Chen, and Jian Wu. Unet 3+: A full-scale connected unet for medical image segmentation, 2020.
- [19] Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. nnU-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, Dec. 2020.
- [20] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.
- [21] Z. Lambert, C. Petitjean, B. Dubray, and S. Ruan. Segthor: Segmentation of thoracic organs at risk in ct images, 2019.
- [22] Bennett Landman, Zhoubing Xu, Juan Eugenio Iglesias, Martin Styner, Thomas Robin Langerak, and Arno Klein. 2015 miccai multi-atlas labeling beyond the cranial vault – workshop and challenge.
- [23] Fangzhou Liao, Ming Liang, Zhe Li, Xiaolin Hu, and Sen Song. Evaluate the malignancy of pulmonary nodules using the 3-d deep leaky noisy-or network. *IEEE transactions on neural networks and learning systems*, 30(11):3484–3495, 2019.
- [24] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, 2015.

- [25] Qing Lyu, Chenyu You, Hongming Shan, and Ge Wang. Super-resolution mri through deep learning. *arXiv preprint arXiv:1810.06776*, 2018.
- [26] Haoyu Ma, Liangjian Chen, Deying Kong, Zhe Wang, Xingwei Liu, Hao Tang, Xiangyi Yan, Yusheng Xie, Shih-Yao Lin, and Xiaohui Xie. Transfusion: Cross-view fusion with transformer for 3d human pose estimation. In *British Machine Vision Conference*, 2021.
- [27] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation, 2016.
- [28] Stanislav Nikolov, Sam Blackwell, Alexei Zverovitch, Ruheena Mendes, Michelle Livne, Jeffrey De Fauw, Yojan Patel, Clemens Meyer, Harry Askham, Bernadino Romera-Paredes, et al. Clinically applicable segmentation of head and neck anatomy for radiotherapy: Deep learning algorithm development and validation study. *Journal of Medical Internet Research*, 23(7):e26151, 2021.
- [29] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.
- [30] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks, 2016.
- [31] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.
- [32] Amber L Simpson, Michela Antonelli, Spyridon Bakas, Michel Bilello, Keyvan Farahani, Bram Van Ginneken, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, et al. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv preprint arXiv:1902.09063*, 2019.
- [33] Shanlin Sun, Yang Liu, Narisu Bai, Hao Tang, Xuming Chen, Qian Huang, Yong Liu, and Xiaohui Xie. Attentionanatomy: A unified framework for whole-body organs at risk segmentation using multiple partially annotated datasets. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2020.
- [34] Hao Tang, Xuming Chen, Yang Liu, Zhipeng Lu, Junhua You, Mingzhou Yang, Shengyu Yao, Guoqi Zhao, Yi Xu, Tingfeng Chen, et al. Clinically applicable deep learning framework for organs at risk delineation in ct images. *Nature Machine Intelligence*, pages 1–12, 2019.
- [35] Hao Tang, Daniel R Kim, and Xiaohui Xie. Automated pulmonary nodule detection using 3d deep convolutional neural networks. In *Biomedical Imaging (ISBI 2018), 2018 IEEE 15th International Symposium on*, pages 523–526. IEEE, 2018.
- [36] Hao Tang, Xingwei Liu, Kun Han, Xiaohui Xie, Xuming Chen, Huang Qian, Yong Liu, Shanlin Sun, and Narisu Bai. Spatial context-aware self-attention model for multi-organ segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 939–949, 2021.
- [37] Hao Tang, Xingwei Liu, Shanlin Sun, Xiangyi Yan, and Xiaohui Xie. Recurrent mask refinement for few-shot medical image segmentation. *arXiv preprint arXiv:2108.00622*, 2021.
- [38] Hao Tang, Xingwei Liu, and Xiaohui Xie. An end-to-end framework for integrated pulmonary nodule detection and false positive reduction. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 859–862. IEEE, 2019.
- [39] Hao Tang, Chupeng Zhang, and Xiaohui Xie. Automatic pulmonary lobe segmentation using deep learning. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 1225–1228. IEEE, 2019.
- [40] Hao Tang, Chupeng Zhang, and Xiaohui Xie. Nodulenet: Decoupled false positive reduction for pulmonary nodule detection and segmentation. *arXiv preprint arXiv:1907.11320*, 2019.
- [41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.
- [42] Zhe Wang, Hao Chen, Xinyu Li, Chunhui Liu, Yuanjun Xiong, Joseph Tighe, and Charless Fowlkes. Sscap: Self-supervised co-occurrence action parsing for unsupervised temporal action segmentation. In *WACV*, 2022.
- [43] Zhe Wang, Liyan Chen, Shauray Rathore, Daeyun Shin, and Charless Fowlkes. Geometric pose affordance: 3d human pose with scene constraints. In *Arxiv*, 2019.
- [44] Zhe Wang, Daeyun Shin, and Charless Fowlkes. Predicting camera viewpoint improves cross-dataset generalization for 3d human pose estimation. In *ECCVW*, 2020.
- [45] Yutong Xie, Jianpeng Zhang, Chunhua Shen, and Yong Xia. Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation. 2021.
- [46] Fan Xu, Haoyu Ma, Junxiao Sun, Rui Wu, Xu Liu, and Youyong Kong. Lstm multi-modal unet for brain tumor segmentation. In *2019 IEEE 4th International Conference on Image, Vision and Computing (ICIVC)*, pages 236–240. IEEE, 2019.
- [47] Chenyu You, Junlin Yang, Julius Chapiro, and James S. Duncan. Unsupervised wasserstein distance guided domain adaptation for 3d multi-domain liver segmentation. In *Interpretable and Annotation-Efficient Learning for Medical Image Computing*, pages 155–163. Springer International Publishing, 2020.
- [48] Chenyu You, Linfeng Yang, Yi Zhang, and Ge Wang. Low-Dose CT via Deep CNN with Skip Connection and Network in Network. In *Developments in X-Ray Tomography XII*, volume 11113, page 111131W. International Society for Optics and Photonics, 2019.
- [49] Chenyu You, Ruihan Zhao, Lawrence Staib, and James S Duncan. Momentum contrastive voxel-wise representation learning for semi-supervised volumetric medical image segmentation. *arXiv preprint arXiv:2105.07059*, 2021.
- [50] Yanci Zhang, Tianming Du, Yujie Sun, Lawrence Donohue, and Rui Dai. Form 10-q itemization. In *Proceedings of*

the 30th ACM International Conference on Information and Knowledge Management (CIKM '21), New York, NY, USA, 2021. Association for Computing Machinery.

- [51] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip H.S. Torr, and Li Zhang. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *CVPR*, 2021.
- [52] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. U-net++: A nested u-net architecture for medical image segmentation, 2018.
- [53] Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: Learning dense volumetric segmentation from sparse annotation, 2016.