Spatial Context-Aware Self-Attention Model For Multi-Organ Segmentation

Hao Tang, Xingwei Liu, Kun Han, Xiaohui Xie
University of California Irvine
{htang6, xingweil, khan7, xhx}@uci.edu

Xuming Chen, Huang Qian, Yong Liu
Shanghai Jiao Tong University
School of Medicine

Shanlin Sun, Narisu Bai
DeepVoxel Inc.

Abstract

Multi-organ segmentation is one of most successful applications of deep learning in medical image analysis. Deep convolutional neural nets (CNNs) have shown great promise in achieving clinically applicable image segmentation performance on CT or MRI images. State-of-the-art CNN segmentation models apply either 2D or 3D convolutions on input images, with pros and cons associated with each method: 2D convolution is fast, less memory-intensive but inadequate for extracting 3D contextual information from volumetric images, while the opposite is true for 3D convolution. To fit a 3D CNN model on CT or MRI images on commodity GPUs, one usually has to either downsample input images or use cropped local regions as inputs, which limits the utility of 3D models for multi-organ segmentation. In this work, we propose a new framework for combining 3D and 2D models, in which the segmentation is realized through high-resolution 2D convolutions, but guided by spatial contextual information extracted from a low-resolution 3D model. We implement a self-attention mechanism to control which 3D features should be used to guide 2D segmentation. Our model is light on memory usage but fully equipped to take 3D contextual information into account. Experiments on multiple organ segmentation datasets demonstrate that by taking advantage of both 2D and 3D models, our method is consistently outperforms existing 2D and 3D models in organ segmentation accuracy, while being able to directly take raw whole-volume image data as inputs.

1. Introduction

Segmentation of organs or lesions from CT images has great clinical implications. It can be used in multiple clinical workflows, including diagnostic interventions, treatment planning and treatment delivery [10]. Organ segmentation is an importance procedure for computer-assisted diagnostic and biomarker measurement systems [38]. Organ-at-risk (OAR) segmentation and tumor segmentation are also crucial to the planning of radiation therapy [32]. Moreover, the segmentation-based models of anatomical structures can support surgical planning and delivery [13].

Organ segmentation is typically done manually by experienced doctors. However, manually segmenting CT image by doctors is often time consuming, tedious and prune to human error, as a typical CT scan can contain up to hundreds of 2D slices. Computational tools that automatically segment organs from CT images can greatly alleviate the doctors' manual effort, given a certain amount of accuracy is achieved.

There is a vast volume of work on organ segmentation using CT or magnetic resonance (MR) image. Traditional segmentation methods are mostly atlas-based. These methods rely on a set of accurate image templates with manual segmentation, and then use image registration to align the new image to the templates. Because of the reliance on the pre-computed templates, these methods 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 [47]. Also, registration is computationally intensive and may take up to hours to complete [4, 7, 5, 12, 16, 24, 31, 35, 36, 43, 41, 14, 9, 3, 40, 42, 6, 50].

Deep learning-based methods provide an alternative solutions with substantial accuracy improvement and speedup. With recent advances in deep learning especially deep convolutional neural network, automatic segmentation using computer algorithm has shown great promise in achieving near human performance [1, 52, 17, 33], and various applications have been deployed in clinical practice.

Fully convolutional network [19] and U-Net [28] are two of the most widely used deep learning-based segmentation algorithms for this purpose. Many its variants have been proposed in recent years, including V-Net [21] and Attention U-Net [23]. These methods can use either 2D or 3D convolutions as its basic building component. 2D methods

usually operate on a slice by slice basis, while 3D methods often operate on a 3D block or a stack of multiple 2D slices [18, 53]. The whole volume prediction can be obtained by predicting each slice or block using a sliding window. Additionally, some may stack multiple 2D slices in the input channel and use 2D convolution as a way to include some 3D features, and this is often referred as 2.5D model.

However, a CT image is inherently 3D. Cutting the images into slices or blocks often ignores the rich information and relation within the whole image volume. A big challenge in developing algorithm for consuming the whole image volume is the GPU memory limitation. Simply storing the tensors of the image features would require huge amount of GPU memory. One way is to adopt a coarse-to-fine strategy [57, 49, 45, 51, 44, 20, 48, 57, 49, 54, 2, 29], where in the first stage the organs of interest are roughly located, and in the second stage the segmentation masks are further refined by using a smaller input based on the localization. This usually requires training multiple CNNs for different stages and organs. Recently, several methods have been proposed to use the whole CT image volume as input, and achieve state-of-the-art accuracy and inference speed [33, 56, 57, 34, 11]. Despite their successes, there exists several disadvantages. First, to reduce the GPU memory consumption, these methods usually directly downsample the input in the very first convolution layer, which may lead to loss of local features. Moreover, they require carefully tailored image input size in order to fit the whole-volume image. However, they will still face GPU memory limitation if the image resolution becomes higher, because the memory requirement grows quickly with the size of the image volume. This makes previous whole-volume algorithm less effective when adapted to new dataset. Second, some of them make strong assumption on the organs/region they segment, thus lacking the ability to generalize well to other parts of the CT image.

We seek to incorporate 3D whole volume information into 2D model in a scalable way. We hypothesize that the benefits of using 3D convolution on the whole-volume image may come from its capability of modeling the shapes and relationships of the 3D anatomical structures. However, to model such shapes and relationships, we do not have to use very high-resolution image. 3D convolution on downsampled image volume may suffice to extract such information and can save a lot of computation and GPU memory. We can use 2D convolution on the original image slice to compensate for the loss of resolution. To fuse both 3D context features and 2D features, we implement a new module called multi-slice feature aggregation based on self-attention [39], which treats the 2D feature map as query and 3D context map as key, and uses self-attention to aggregate the rich 3D context information.

In this paper, we propose a new deep learning framework named **S**patial **C**ontext-**A**ware Self-**A**ttention Model (SCAA). Our main contributions are: i) a new framework for combining 3D and 2D models that takes the whole-volume CT image as input; ii) a self-attention mechanism to filter and aggregate 3D context features from the whole volume image to guide 2D segmentation iii) the proposed method can scale to larger input volume without concerning the GPU memory limitation that common 3D methods face. Experiment results on a head and neck (HaN) dataset of 9 organs and an abdomen dataset of 11 organs show the proposed model consistently outperforms state-of-the-art methods in terms of organ segmentation accuracy, while being able to take the whole-volume CT image as input.

2. Method

Figure 1 shows the details of the proposed method. The proposed model consists of four parts: a 3D context feature encoder f^{3D}, a 2D feature encoder f^{Enc}, a multi-slice feature aggregation (MSFA) module f^{MSFA}, and at last a 2D decoder f^{Dec}. The input to the model $f : x \rightarrow f^{Dec}(f^{Enc}(f^{3D}(x)))$ is the whole CT image volume $\mathbf{I} \in \mathbb{R}^{D\times H\times W}$, and the outputs are D 2D segmentation masks for C classes $\mathbf{m} \in \mathbb{R}^{D\times H\times W\times C}$. D, H, W are the depth, height and width of the image volume.

2.1. 3D context feature encoder and 2D encoder

f^{3D} first downsamples the input 3D volume $\mathbf{I}$ to $\mathbf{I}' \in \mathbb{R}^{D^{3D}\times H^{3D}\times W^{3D}}$. D^{3D}, H^{3D}, W^{3D} are the depth, height, and width of the downsampled 3D volume. We use a downsample factor of two in this work. Note we can also downsample the volume to other resolutions, e.g. isotropic 4mm resolution. It then applies 3D convolution blocks three times, where each convolution block consists of two residual blocks followed by one $2 \times 2 \times 2$ max pooling, aiming at extracting 3D context features in the whole CT image.

The output of f^{3D} are four feature maps at different scales, denoted as $F_i^{3D} \in \mathbb{R}^{C_i^{3D}\times D_i^{3D}\times H_i^{3D}\times W_i^{3D}}$, where $i = 2, 3, 4, 5$. This means the feature map F_i^{3D} is downsampled by a factor of 2^i compared to the original image. D_i^{3D}, H_i^{3D} and W_i^{3D} are the depth, height and width of the feature map at scale i, the values of which depend on the size of input image. C_i^{3D} equals 24, 32, 64 and 64 for $i = 2, 3, 4, 5$ respectively. After F_5^{3D}, we flatten the channel, depth, height and width dimension into a vector and regard it as a global descriptor F_{globe} for the 3D volume.

f^{Enc} is similar to U-Net encoder. It consumes one axial slice of the CT image $\mathbf{S} \in \mathbb{R}^{H\times W}$ and applies 2D convolution blocks five times, where each block consists of two convolutions followed by instance normalization and ReLU activation, and a max pooling at the end. The 2D feature encoder outputs five sets of feature maps at different scales, denoted as $F_i^{2D} \in \mathbb{R}^{C_i^{3D}\times H_i^{2D}\times W_i^{2D}}$, where

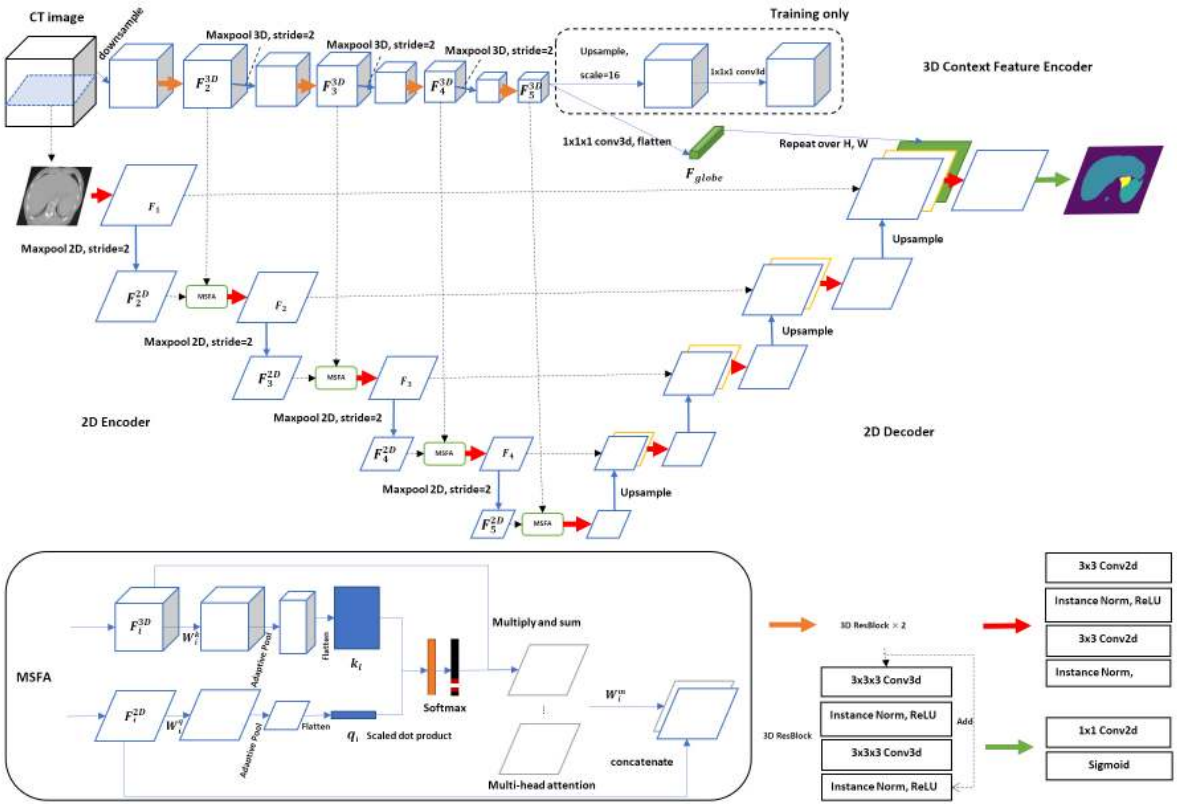

Figure 1: **Overview of spatial context-aware self-attention model (SCAA)**. SCAA consists of a 3D context feature encoder, a 2D encoder, a 2D decoder and a multi-slice feature aggregation (MSFA) module. SCAA starts with extracting 3D features from the downsampled CT image using 3D convolutions. Then the 2D encoder extracts 2D features and uses MSFA module to fuse 2D and 3D features hierarchically. Lastly, the 2D decoder decodes the fused 2D and 3D features and outputs 2D segmentation masks of each organ. The numbers of feature channels in F_i are 64, 96, 128, 192, 256 for $i = 1, 2, 3, 4, 5$ respectively. The number of feature channels in F_i^{2D} is 96, 128, 192, 256 for $i = 2, 3, 4, 5$ respectively. The numbers of feature channels in F_i^{3D} are 24, 32, 64, 64 for $i = 2, 3, 4, 5$ respectively. W_i^k is implemented by using $1 \times 1 \times 1$ 3D convolution with 2, 2, 4, and 4 feature channels for $i = 2, 3, 4, 5$ respectively. W_i^q is implemented by using 1×1 2D convolution with 2, 2, 4, and 4 feature channels for $i = 2, 3, 4, 5$ respectively. The xy spatial resolution of the output of the adaptive pooling are 16×16, 8×8, 4×4, 4×4 for scale $i = 2, 3, 4, 5$ respectively. The number of attention heads for scale $i = 2, 3, 4, 5$ is 2, 2, 4, and 4 respectively.

$i = 1, 2, 3, 4, 5$. C_i^{2D} equals 64, 96, 128, 192 and 256 for $i = 2, 3, 4, 5$ respectively.

2.2. Multi-scale feature aggregation

Inspired by Transformer [39, 25], we implement a self-attention mechanism to filter and extract useful 3D context features from our 3D feature maps F_i^{3D}, and we name this module as multi-scale feature aggregation (MSFA). We regard 3D features as values, and generate queries from the 2D features and keys from the 3D features. Based on feature similarities of current 2D features and all slices in the 3D feature map (along z dimension), the MSFA will generate an attention vector $\mathbf{a_i} \in \mathbb{R}^{D_i^{3D}}$ the same length of the depth of the 3D feature map. This attention then is applied to the 3D feature map to generate a 2D feature map that is considered as the aggregated 3D context features.

We start by mapping our 2D feature map F_i^{2D} and 3D feature map F_i^{3D} ($i = 2, 3, 4, 5$), to one query and D_i^{3D} keys, of the same embedding space. We use a weight metric W_i^q (1×1 2D convolution) to generate our query $q \in \mathbb{R}^{C_{embed} \times H_i \times W_i}$. We use a weight metric W_i^k ($1 \times 1 \times 1$ 3D convolution) to generate our keys $\{k_j\}$ of size $C_i^{embed} \times H_i^{3D} \times W_i^{3D}$, where $j = 1, 2, ..., D_i^{3D}$. An adaptive pooling operation is used to reduce the spatial resolution of the query and keys to $H_i' \times W_i'$, followed by a flatten operation to make them one dimensional. As a result, the embed dimension of the query and keys is now of size

$C_i^{embed} \cdot H_i' \cdot W_i'$:

$$
\begin{aligned}
(F_i^{3D'})_{c'dhw} &= (F_i^{3D})_{cdhw}(W_i^k)_{c'c} \\
(F_i^{2D'})_{c'hw} &= (F_i^{2D})_{chw}(W_i^q)_{c'c} \\
k_i &= \text{Flatten}(\text{AdaptivePool}_{D_i^{3D}H_i'W_i'}(F_i^{3D'})) \\
q_i &= \text{Flatten}(\text{AdaptivePool}_{H_i'W_i'}(F_i^{2D'},))
\end{aligned} \tag{1}
$$

$(\cdot)_{(\cdot)}$ is the Einstein summation convention.

A scaled dot product is used to compute the response of the 2D feature map with the 3D feature map $r_j = \frac{q \cdot k_j}{\sqrt{C_i^{embed} \cdot H_i' \cdot W_i'}}$. A softmax is followed to generate our attention $\mathbf{a_i} = \text{softmax}(\mathbf{r})$. We then multiply the attention $\mathbf{a_i}$ over the depth dimension of F_i^{3D} and sum over the depth dimension to generate our aggregated context feature map $F_i^{agg} \in \mathbb{R}^{C_i^{3D} \times H_i \times W_i}$:

$$
(F_i^{agg})_{chw} = (F_i^{3D})_{cdhw}(\mathbf{a_i})_d \tag{2}
$$

A multi-head attention mechanism is also used. We generate m_i such fused 2D feature map and then use a weight metric W_i^m (1×1 2D convolution) to aggregate multiple self-attention output. m_i is 2, 2, 4 and 4 for $i = 2, 3, 4, 5$ respectively. This multi-head attention allows our model to focus on different parts of the 3D context volume to extract features required by different classes. Two 2D convolution blocks on the concatenated 2D feature map of F_i^{2D} and F_i^{agg} are used to better combine the 2D and 3D context features. F_i denotes our final 2D feature map for scale i. Note that F_1 is the same as F_1^{2D}.

2.3. 2D decoder

f^{Dec} is similar to the U-Net decoder. Starting from F_5, a 2D upsample is used first to increase the spatial resolution by 2. Then we concatenate the upsampled features with the corresponding encoder feature map and apply one 2D convolution block. The last upsampled feature map is of the same resolution as our input image. We concatenate our 3D global descriptor F_{globe} to each pixel's feature vector and use a 1×1 convolution to obtain the final axial segmentation mask for each class $\{\mathbf{m}_c^{2D}\}$, where $c \in \mathbb{Z}_{<C}^*$ and C is the number of classes.

2.4. Loss function and implementation details

The loss function is defined as:

$$
L^{2D} = \sum_c^N 1 - \phi(\mathbf{m}^c, \mathbf{g}^c) \tag{3}
$$

$\mathbf{g}$ is the ground truth segmentation for the axial slice. $\phi(\mathbf{m}, \mathbf{g})$ computes a soft Dice score between the predicted mask $\mathbf{m}$ and the ground truth $\mathbf{g}$:

$$
\phi(\mathbf{m}, \mathbf{g}) = \frac{\sum_i^N \mathbf{m}_i \mathbf{g}_i}{\sum_i^N \mathbf{m}_i \mathbf{g}_i + \alpha \sum_i^N \mathbf{m}_i (1 - \mathbf{g}_i) + \beta \sum_i^N (1 - \mathbf{m}_i) \mathbf{g}_i + \epsilon} \tag{4}
$$

N is the number of total pixels in the batch. α and β are two hyper parameters controlling the penalty for false positive and true negative respectively, and we set them to both 0.5. ϵ is used for numerical stability.

To facilitate the training of 3D context feature encoder, we add an auxiliary 3D segmentation loss. We first upsample F_5^{3D} by a factor of 16, so it has the same spatial resolution as the downsampled 3D image volume. A $1 \times 1 \times 1$ 3D convolution is used to obtain the 3D segmentation mask $\{m_c^{3D}\}$. We use the same dice loss to get our 3D supervision loss L^{3D}. The final loss is then $L = L^{3D} + L^{2D}$.

We use one CT image for each batch during training. For each batch, we generate one 3D image volume and randomly sample 16 axial slices (batch size 16 for the 2D network). We only need to forward the 3D context encoder once per batch. We use Adam with initial learning rate 10^{-4} as optimizer for a total of 150 epochs. We applied elastic transformation and random jitter for data augmentation.

For testing, we only need to forward the 3D context feature encoder once, and we segment each 2D slice using the 2D decoder/encoder and MSFA.

3. Experiments

3.1. Datasets

Two datasets were used for evaluation: i) MICCAI 2015 head and neck (HaN)organ-at-risk (OAR) segmentation challenge dataset [26], containing a training of 33 CT images and a test of 15 CT images. The dataset contains manually labeled organ segmentation mask for 9 organs: brain stem, mandible, optic nerve left and right, optic chiasm, parotid left and right, submandibular gland (SMG) left and right; ii) an in-house abdomen multi-organ segmentation dataset[1] (ABD-110) containing 110 contrast enhanced CT images and 11 organs (large bowel, duodenum, spinal cord, liver, spleen, small bowel, pancreas, left and right kidney, stomach and gallbladder). The 110 CT scans were collected from 110 patients who had radiotherapy during the past three years. The CT scans were manually delineated by one experienced doctor and then manifested by another. We use the official split of training set to train the model and test on the official test set on MICCAI 2015 challenge dataset, following the same protocol as previous work [33, 22, 26, 11]. All experiments on ABD-110 dataset was conducted using 4-fold cross validation.

We report the segmentation performance using dice similarity coefficient (DSC) in percentage and 95% hausdorff

[1]Use of this dataset has been approved by an institutional review board (IRB).

distance (HD) in mm following previous work [26]. DSC measures the overlap between the predicted mask $\mathbf{m}$ and ground truth mask $\mathbf{g}$:

$$\mathbf{DSC} = \frac{2|\mathbf{m} \cup \mathbf{g}|}{|\mathbf{m} \cap \mathbf{g}|} \tag{5}$$

3.2. Ablation study on ABD-110

To compare different ways of integrating 3D features and demonstrate the contribution of each of the add-on modules in the proposed model, we conducted ablation studies with the following different settings: 1) model with only a 3D global descriptor F_{globe} concatenated to the last feature map (CA), to demonstrate the importance of including 3D features progressively during feature extraction. 2) model without MSFA module, but only uses the corresponding center slice feature from the 3D context feature maps (C-CA), to show the effectiveness and importance of using self-attention to aggregate 3D features. 3) SCAA model without concatenating F_{globe} to the last feature map (SCAA*), to show whether F_{globe} is crucial to improve the segmentation accuracy.

As we can see from Table 1, by adding the 3D global descriptor to the 2D U-Net, CA outperforms the 2D U-Net by 0.4% and lowers the 95%HD by 0.7 mm, showing the importance of integrating the 3D holistic information. Next, by progressively integrating 3D features to 2D feature extractor, C-CA outperforms CA by 1.4% in DSC and 0.2 mm in 95%HD, showing the importance of integrating 3D context features hierarchically. Finally, by adding the MSFA module based on self-attention, SCAA outperforms C-CA by 0.4%, demonstrating the effectiveness of the implemented self-attention mechanism to weigh the information from different adjacent slices. To find out which part of the 3D features contribute the most, we then compare the performance of SCAA and SCAA*. SCAA* achieves average DSC of 86.9% and average 95%HD 3.6 mm, while SCAA achieves 86.8% and 4.3 mm. They are very close in terms of DSC and SCAA* wins on 95%HD. This demonstrates that F_{globe} has very little contribution, and progressively integrating 3D features with 2D encoder achieves the most performance gain.

For all following comparisons with other methods, we use the best performing configuration SCAA*.

3.3. Comparison with previous methods on ABD-110

To compare with previous methods of multi-organ segmentation on the ABD-110 dataset, we ran the following representative algorithms: U-Net [28], Attention U-Net [23], U_a-Net [33], and nnU-Net [15]. U-Net is a well-established medical image segmentation baseline algorithm. Attention U-Net is a multi-organ segmentation framework that uses gated attention to filter out irrelevant response in the feature maps. U_a-Net is a state-of-the-art end-to-end two-stage framework for multi-organ segmentation in the head and neck region. nnU-Net is a self-adaptive medical image semantic segmentation framework that wins the first in the Medical Segmentation Decathlon (MSD) challenge [30]. nnU-Net mainly consists of three main deep learning-based segmentation methods: a 2D U-Net (slice-wise), a 3D U-Net (patch-wise) and a coarse-to-fine cascade framework consisting of two 3D U-Nets. Its final model is an ensemble of the three methods. The above-mentioned works cover a wide range of algorithms for multi-organ segmentation and should provide a comprehensive and fair comparison to our proposed method on the in-house dataset. For 3D Attention U-Net, we followed the same preprocessing as in its original paper, to downsample the image to isotropic 2mm resolution due to GPU memory limitation. However, for all other methods, we feed the original CT image with its original image spacing.

The results are shown in section 3.3. First, by comparing 2D and 3D methods, we can see that the performance of 2D methods is on par with 3D methods on kidneys, spinal cord and liver, which is likely because those organs are usually large and have regular shapes. However, for organs like stomach, small and large bowels, 3D methods generally perform better. This may be because those organs often have more anatomical variance, and a 3D holistic understanding of the context is beneficial. Next, U_a-Net was 1.4% lower than 2D U-Net and 3.5% lower than SCAA. This may be because U_a-Net was designed mainly for the head and neck region where organs are relatively small and do not overlap too much with each other. The abdomen region, on the other hand, is more complicated as the bounding boxes of some organs overlap a lot with each other (e.g. large bowel and small bowel), which makes U_a-Net less effective. Finally, comparing SCAA to nnU-Net, we find SCAA outperform nnU-Net by 1.2%. The best configuration of nnU-Net on ABD-110 was ensemble of a 2D U-Net (slice by slice) and a 3D U-Net (patch-wise). Both nnU-Net and SCAA consider the fusion of 2D model and 3D model, but they implement it in different ways - nnU-Net uses ensemble to combine 2D and 3D models while SCAA integrates 3D model into 2D model in an end-to-end fashion and jointly optimizes both models. This improvement then is likely due to the soft attention mechanism that allows SCAA to filter and extract relevant features from the large 3D context and better fuse the 2D and 3D models. Altogether, we demonstrated the effectiveness of the proposed method, which achieves an average DSC of 86.9% on the in-house dataset.

3.4. Performance on MICCAI2015

A second multi-organ segmentation dataset from MICCAI 2015 organ-at-risk (OAR) segmentation challenge[26]

Anatomy	U-Net	CA	C-CA	SCAA	SCAA*
Large Bowel	80.5 ± 9.4	79.6 ± 10.2	81.5 ± 9.0	81.5 ± 10.0	$\mathbf{82.5 \pm 9.2}$
Duodenum	63.4 ± 18.6	67.6 ± 17.4	69.9 ± 17.2	$\mathbf{71.4 \pm 17.2}$	70.7 ± 17.5
Spinal Cord	90.3 ± 3.8	90.4 ± 3.8	$\mathbf{91.0 \pm 3.7}$	90.7 ± 4.0	90.8 ± 3.5
Liver	95.5 ± 1.9	96.0 ± 1.9	96.2 ± 1.4	96.4 ± 1.1	$\mathbf{96.4 \pm 1.2}$
Spleen	94.6 ± 3.1	95.2 ± 2.3	95.4 ± 2.0	95.6 ± 2.3	$\mathbf{95.9 \pm 1.4}$
Small Bowel	72.2 ± 16.2	72.4 ± 16.0	75.4 ± 16.0	76.1 ± 15.1	$\mathbf{76.5 \pm 15.3}$
Pancreas	79.8 ± 9.1	79.9 ± 10.6	81.8 ± 9.4	81.8 ± 8.5	$\mathbf{82.1 \pm 9.1}$
Kidney L	95.7 ± 1.2	95.7 ± 1.8	95.8 ± 1.4	$\mathbf{96.0 \pm 1.4}$	96.0 ± 1.5
Kidney R	95.3 ± 3.0	95.5 ± 2.8	95.6 ± 3.3	95.6 ± 2.7	$\mathbf{95.7 \pm 2.5}$
Stomach	84.2 ± 16.7	85.0 ± 15.8	86.1 ± 15.6	86.8 ± 13.6	$\mathbf{87.5 \pm 14.3}$
Gallbladder	78.6 ± 19.5	78.2 ± 20.1	81.4 ± 18.1	$\mathbf{82.7 \pm 17.2}$	82.2 ± 17.7
Average	84.6	85.0	86.4	86.8	**86.9**

Anatomy	U-Net	CA	C-CA	SCAA	SCAA*
Large Bowel	9.5 ± 7.8	9.7 ± 8.3	8.9 ± 8.7	7.1 ± 4.6	$\mathbf{6.6 \pm 5.0}$
Duodenum	7.8 ± 4.9	7.4 ± 4.9	6.6 ± 4.7	6.2 ± 4.8	$\mathbf{5.7 \pm 4.1}$
Spinal Cord	1.8 ± 2.6	1.9 ± 2.8	1.8 ± 2.5	1.9 ± 3.0	$\mathbf{1.6 \pm 2.3}$
Liver	3.9 ± 3.9	2.5 ± 2.6	2.6 ± 3.4	2.1 ± 1.5	$\mathbf{1.9 \pm 1.4}$
Spleen	6.5 ± 12.8	2.4 ± 4.7	2.5 ± 7.3	1.7 ± 4.6	$\mathbf{1.2 \pm 0.7}$
Small Bowel	7.8 ± 7.3	8.1 ± 7.8	9.0 ± 11.4	8.3 ± 8.4	$\mathbf{7.4 \pm 7.1}$
Pancreas	4.1 ± 3.3	3.9 ± 3.9	3.6 ± 3.4	3.5 ± 3.5	$\mathbf{3.3 \pm 3.7}$
Kidney L	1.5 ± 1.4	1.5 ± 1.1	1.9 ± 5.4	1.2 ± 0.6	$\mathbf{1.2 \pm 0.4}$
Kidney R	1.8 ± 3.2	1.5 ± 1.6	$\mathbf{1.3 \pm 1.0}$	1.7 ± 2.4	1.3 ± 1.1
Stomach	7.2 ± 8.1	6.6 ± 7.4	$\mathbf{5.7 \pm 7.4}$	8.2 ± 10.5	5.9 ± 7.9
Gallbladder	7.0 ± 11.5	6.0 ± 7.6	6.0 ± 11.4	4.9 ± 9.1	$\mathbf{3.1 \pm 4.6}$
Average	5.4	4.7	4.5	4.3	**3.6**

Table 1: **Ablation study on different ways of fusing 2D and 3D features. Left:** DSC (unit: %). Higher the better. **Right:** 95%HD (unit: mm). Lower the better. Bold numbers represent the best performance. CA stands for context-aware model, which does not progressively integrate 3D features from the 3D model. C-CA for center context-aware model, which only integrates the corresponding center slice from the 3D feature maps. SCAA for spatial context-aware self-attention model, which uses the MSFA module to aggregate 3D features from the whole 3D volume. SCAA* for model without concatenating F_{globe} to the last feature map.

Organ	U-Net[1]	Att. U-Net[1]	Att. U-Net	U-Net (3D patch)[2]	U_a-Net	nnU-Net	SCAA
Large Bowel	80.5 ± 9.4	80.3 ± 9.3	80.2 ± 8.7	80.7 ± 9.7	77.1 ± 10.4	82.1 ± 8.5	$\mathbf{82.5 \pm 9.2}$
Duodenum	63.4 ± 18.6	64.5 ± 18.1	67.1 ± 16.0	70.2 ± 16.3	62.6 ± 17.7	$\mathbf{71.3 \pm 15.8}$	70.7 ± 17.5
Spinal Cord	90.3 ± 3.8	90.9 ± 4.0	89.8 ± 3.9	88.4 ± 4.6	$\mathbf{91.6 \pm 3.5}$	89.5 ± 4.6	90.8 ± 3.5
Liver	95.5 ± 1.9	95.7 ± 1.5	96.0 ± 1.1	95.9 ± 1.2	94.7 ± 1.9	96.4 ± 1.0	$\mathbf{96.4 \pm 1.2}$
Spleen	94.6 ± 3.1	95.1 ± 2.8	95.0 ± 1.9	92.9 ± 13.2	94.7 ± 2.1	93.8 ± 12.7	$\mathbf{95.9 \pm 1.4}$
Small Bowel	72.2 ± 16.2	73.7 ± 16.6	75.0 ± 13.8	73.7 ± 15.6	75.0 ± 13.8	75.7 ± 14.3	$\mathbf{76.5 \pm 15.3}$
Pancreas	79.8 ± 9.1	80.4 ± 9.8	79.0 ± 10.0	80.2 ± 12.1	76.3 ± 11.6	81.8 ± 11.5	$\mathbf{82.1 \pm 9.1}$
Kidney L	95.7 ± 1.2	95.6 ± 1.6	94.5 ± 9.0	94.2 ± 9.0	95.2 ± 1.5	94.8 ± 9.1	$\mathbf{96.0 \pm 1.5}$
Kidney R	95.3 ± 3.0	95.5 ± 2.9	94.9 ± 3.1	94.2 ± 9.2	94.8 ± 2.5	94.5 ± 9.2	$\mathbf{95.7 \pm 2.5}$
Stomach	84.2 ± 16.7	84.0 ± 15.2	85.4 ± 15.6	87.3 ± 14.5	83.1 ± 14.1	$\mathbf{88.0 \pm 16.0}$	87.5 ± 14.3
Gallbladder	78.6 ± 19.5	77.1 ± 20.5	78.2 ± 19.1	74.2 ± 29.6	72.6 ± 26.8	75.0 ± 29.7	$\mathbf{82.2 \pm 17.7}$
Average	84.6	84.8	85.0	84.7	83.4	85.7	**86.9**

[1] 2D U-Net/Attention U-Net with adjacent 3 slices stacked into channel as input (2.5D).
[2] 3D U-Net with patch-wise input.

Table 2: **Dice similarity coefficient (DSC) comparison on ABD-110.** Bold number means the best DSC performance. SCAA (proposed) achieves the highest DSC compared to other methods.

was used for evaluation. First, as we can see from Table 3, SCAA outperforms [22] by 4.2%. [22] used a combination of 3D and 2D convolution on 21 stacked slices for OAR segmentation. This shows the use of larger context information is beneficial for a good segmentation accuracy. Next, by comparing SCAA to AnatomyNet [55] which is a 3D model that takes the whole-volume CT as input, SCAA was 2.4% higher. This is likely due to the attention mechanism that helps the model to filter irrelevant features from the entire volume. Also, SCAA outperforms U_a-Net [33] by 1.4%. U_a-Net is an end-to-end two-stage model that first detects bounding box of OARs and then segments organs within the bounding box. SCAA performed better, partly because SCAA did not enforce a 'hard' attention (bounding box) but rather use 'soft' attention to enable the model focus on a smaller region. This keeps SCAA away from potential bounding box regression error and missing detection. Finally, SCAA outperforms previous state-of-the-art method [11] in 5 out of 9 organs and achieves an average DSC of 82.6%, 0.2% higher than the state-of-the-art method. Also note that [11] is a two-stage segmentation framework, which consists of two DCNNs. Our proposed method (SCAA), however, is a one-stage end-to-end solution for multi-organ segmentation, requiring less training time and computation, as well as fewer parameters.

3.5. Memory consumption

One advantage of the proposed method is that it significantly reduces the GPU memory while at the same time preserves the large 3D context features. To demonstrate GPU memory consumption when using whole-volume as input, we estimated and measured the actual GPU memory cost (using PyTorch as framework) for different 3D models during training in Table 4. We made several assumptions: i) the input image volume is of size $256 \times 256 \times 256$. This is the size used for whole-volume input with original image spac-

Study	Brain Stem	Mandible	Optic Chiasm	Optic Nerve		Parotid		SMG		Avg.
				L	R	L	R	L	R	
Raudashl *et al*. [26]	88.0	93.0	55.0	62.0	62.0	84.0	84.0	78.0	78.0	76.0
Fritscher *et al*. [8]			49.0 ± 9.0			81.0 ± 4.0	81.0 ± 4.0	65.0 ± 8.0	65.0 ± 8.0	-
Ren *et al*. [27]			58.0 ± 17.0	72.0 ± 8.0	70.0 ± 9.0					-
Wang *et al*. [46]	**90.0 ± 4.0**	94.0 ± 1.0				83.0 ± 6.0	83.0 ± 6.0			-
Zhu *et al*. [55]	86.7 ± 2.0	92.5 ± 2.0	53.2 ± 15.0	72.1 ± 6.0	70.6 ± 10.0	88.1 ± 2.0	87.3 ± 4.0	81.4 ± 4.0	81.3 ± 4.0	79.2
Tong *et al*. [37]	87.0 ± 3.0	93.7 ± 1.2	58.4 ± 10.3	65.3 ± 5.8	68.9 ± 4.7	83.5 ± 2.3	83.2 ± 1.4	75.5 ± 6.5	81.3 ± 6.5	77.4
Nikolov *et al*. [22]	79.5 ± 7.8	94.0 ± 2.0		71.6 ± 5.8	69.7 ± 7.1	86.7 ± 2.8	85.3 ± 6.2	76.0 ± 8.9	77.9 ± 7.4	-
Tang *et al*. [33]	87.5 ± 2.5	95.0 ± 0.8	61.5 ± 10.2	74.8 ± 7.1	72.3 ± 5.9	88.7 ± 1.9	87.5 ± 5.0	82.3 ± 5.2	81.5 ± 4.5	81.2
Guo *et al*. [11]	87.6 ± 2.8	95.1 ± 1.1	**64.5 ± 8.8**	75.3 ± 7.1	74.6 ± 5.2	88.2 ± 3.2	88.2 ± 5.2	**84.2 ± 7.3**	**83.8 ± 6.9**	82.4
SCAA (proposed)	89.2 ± 2.6	**95.2 ± 1.3**	62.0 ± 16.9	**78.4 ± 6.1**	**76.0 ± 7.5**	**89.3 ± 1.5**	**89.2 ± 2.3**	83.2 ± 4.9	80.7 ± 5.2	**82.6**

Table 3: **Comparison of DSC with previous methods on the MICCAI 2015 9 organs segmentation challenge.** Numbers are the higher the better (best in bold).

Method	Batch size	Estimate (GB)	Actual (GB)	# of parameters
2D U-Net [28]	4	2.86	3.35	34.51 M
3D U-Net [23]	1	27.96	out of memory	5.88 M
3D Attention U-Net [23]	1	17.31	out of memory	6.40 M
SCAA (3D encoder)	1	3.22	-	-
SCAA (2D U-Net & MSFA)	4	2.13	-	-
SCAA (total)		5.35	6.44	7.82 M

Table 4: **GPU memory consumption comparison using whole-volume image as input of, and number of parameters for different methods on ABD-110.** We used PyTorch as the deep learning framework to measure the actual GPU memory cost.

ing. ii) we only take the memory cost of storing tensor and its gradient after each convolution and batch/instance normalization layer into consideration, because they consume the most GPU memory. iii) each number is a floating point number (32 bits). For 2D U-Net, the numbers of channel for the five scales are 64, 128, 256, 512 and 1024 respectively, as in the original implementation [28]. For the 3D U-Net, as the network has more parameters in convolution kernels, fewer channels are used in practice (16, 32, 64, 128 and 256) [23]. The memory cost and number of parameters of [28] and [23] were computed by running their released code. The GPU memory cost is for the training phase, and that of inference is approximately half of the values in Table 4. The actual cost is computed by running the algorithm on a GTX 1080 Ti GPU card (12 GB memory).

As seen from Table 4, compared to most 3D U-Net based methods, we only require 6.44 GB total memory for a batch size of four during training, which is approximately 35.1% of the 3D Attention U-Net, demonstrating the efficiency of the proposed method. Moreover, our method supports distributing batches among multiple GPU devices, which is more scalable than previous 3D methods.

3.6. Visualization

We visualize the attention vector $\mathbf{a_i}$ learnt for F_i^{3D} Figure 2 and the prediction of a random CT image from the ABD-110 dataset Figure 3. As seen from Figure 2, the 3D slice features that are useful when segmenting each 2D slice are mostly its adjacent slices. This accords with the intuition that the most prominent and useful 3D information should be mostly from its neighbouring slices. But it is also important to incorporate full 3D context information progressively. We have demonstrated the effectiveness of the self-attention mechanism in Section 3.2 by comparing to C-CA that only integrates the corresponding center slice

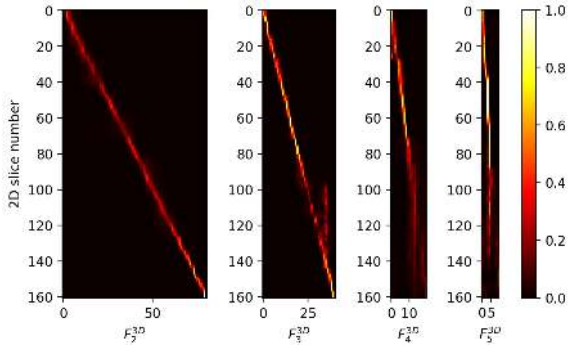

Figure 2: Attention vector learnt by the proposed method. Y-axis is the slice number of the CT image, and the X-axis is the slice number (depth) of the 3D feature map F_i^{2D}.

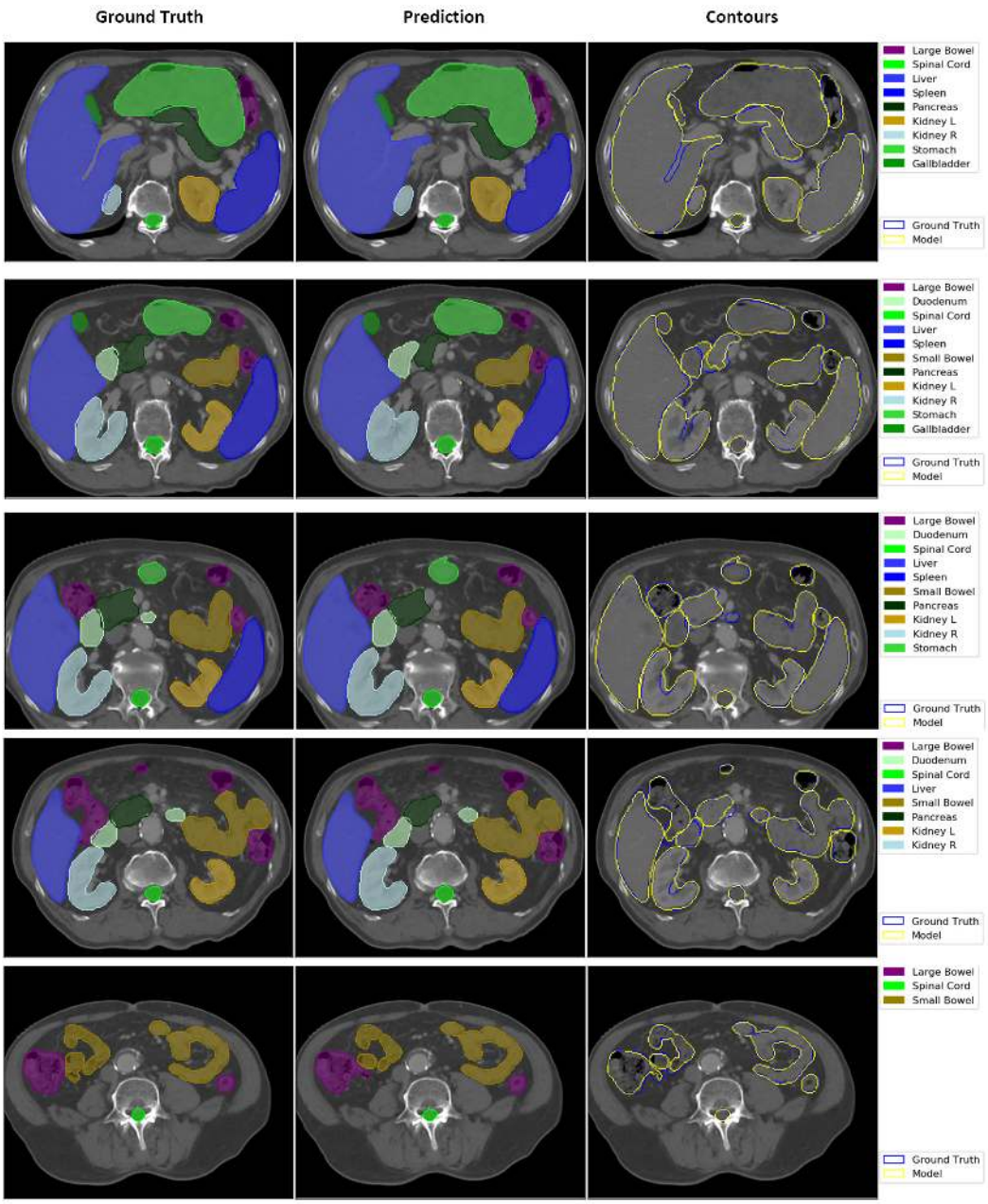

Figure 3: **A CT image from ABD-110 dataset**. The first and second columns are the ground truth and predicted mask overlayed on the original CT image slice, respectively. The third column shows the comparison of contours of the ground truth and predicted mask on the same slice.

feature from the 3D feature map.

4. Conclusion

In this paper, we propose a Spatial Context-aware Self-Attention model for multi-organ segmentation. The proposed model uses a self-attention mechanism to filter useful 3D contextual information from the large 3D whole-volume CT image to guide the segmentation of 2D slice. It addresses the GPU memory concerns that common whole volume-based 3D methods confront. Experiments on two multi-organ segmentation datasets demonstrate the state-of-the-art performance of the proposed model.

References

[1] Wenjia Bai, Matthew Sinclair, Giacomo Tarroni, Ozan Oktay, Martin Rajchl, Ghislain Vaillant, Aaron M Lee, Nay Aung, Elena Lukaschuk, Mihir M Sanghvi, et al. Human-level cmr image analysis with deep fully convolutional networks. *arXiv preprint arXiv:1710.09289*, 2017.

[2] Jinzheng Cai, Le Lu, Yuanpu Xie, Fuyong Xing, and Lin Yang. Improving deep pancreas segmentation in ct and mri images via recurrent neural contextual learning and direct loss function. *arXiv preprint arXiv:1707.04912*, 2017.

[3] Olivier Commowick, Vincent Grégoire, and Grégoire Malandain. Atlas-based delineation of lymph node levels in head and neck computed tomography images. *Radiotherapy and Oncology*, 87(2):281–289, 2008.

[4] Jean-François Daisne and Andreas Blumhofer. Atlas-based automatic segmentation of head and neck organs at risk and nodal target volumes: a clinical validation. *Radiation oncology*, 8(1):154, 2013.

[5] Hoang Duc, K Albert, Gemma Eminowicz, Ruheena Mendes, Swee-Ling Wong, Jamie McClelland, Marc Modat, M Jorge Cardoso, Alex F Mendelson, Catarina Veiga, et al. Validation of clinical acceptability of an atlas-based segmentation algorithm for the delineation of organs at risk in head and neck cancer. *Medical physics*, 42(9):5027–5034, 2015.

[6] Valerio Fortunati, René F Verhaart, Wiro J Niessen, Jifke F Veenland, Margarethus M Paulides, and Theo van Walsum. Automatic tissue segmentation of head and neck mr images for hyperthermia treatment planning. *Physics in Medicine & Biology*, 60(16):6547, 2015.

[7] Valerio Fortunati, René F Verhaart, Fedde van der Lijn, Wiro J Niessen, Jifke F Veenland, Margarethus M Paulides, and Theo van Walsum. Tissue segmentation of head and neck ct images for treatment planning: a multiatlas approach combined with intensity modeling. *Medical physics*, 40(7), 2013.

[8] Karl Fritscher, Patrik Raudaschl, Paolo Zaffino, Maria Francesca Spadea, Gregory C Sharp, and Rainer Schubert. Deep neural networks for fast segmentation of 3d medical images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 158–165. Springer, 2016.

[9] Karl D Fritscher, Marta Peroni, Paolo Zaffino, Maria Francesca Spadea, Rainer Schubert, and Gregory Sharp. Automatic segmentation of head and neck ct images for radiotherapy treatment planning using multiple atlases, statistical appearance models, and geodesic active contours. *Medical physics*, 41(5), 2014.

[10] 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.

[11] Dazhou Guo, Dakai Jin, Zhuotun Zhu, Tsung-Ying Ho, Adam P Harrison, Chun-Hung Chao, Jing Xiao, and Le Lu. Organ at risk segmentation for head and neck cancer using stratified learning and neural architecture search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4223–4232, 2020.

[12] MS Hoogeman, X Han, D Teguh, P Voet, P Nowak, T Wolf, L Hibbard, B Heijmen, and P Levendag. Atlas-based auto-segmentation of ct images in head and neck cancer: What is the best approach? *International Journal of Radiation Oncology• Biology• Physics*, 72(1):S591, 2008.

[13] Robert D Howe and Yoky Matsuoka. Robotics for surgery. *Annual review of biomedical engineering*, 1(1):211–240, 1999.

[14] Aurélie Isambert, Frédéric Dhermain, François Bidault, Olivier Commowick, Pierre-Yves Bondiau, Grégoire Malandain, and Dimitri Lefkopoulos. Evaluation of an atlas-based automatic segmentation software for the delineation of brain organs at risk in a radiation therapy clinical context. *Radiotherapy and oncology*, 87(1):93–99, 2008.

[15] Fabian Isensee, Jens Petersen, Andre Klein, David Zimmerer, Paul F Jaeger, Simon Kohl, Jakob Wasserthal, Gregor Koehler, Tobias Norajitra, Sebastian Wirkert, et al. nnu-net: Self-adapting framework for u-net-based medical image segmentation. *arXiv preprint arXiv:1809.10486*, 2018.

[16] PC Levendag, M Hoogeman, D Teguh, T Wolf, L Hibbard, O Wijers, B Heijmen, P Nowak, E Vasquez-Osorio, and X Han. Atlas based auto-segmentation of ct images: Clinical evaluation of using auto-contouring in high-dose, high-precision radiotherapy of cancer in the head and neck. *International Journal of Radiation Oncology• Biology• Physics*, 72(1):S401, 2008.

[17] 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.

[18] Siqi Liu, Daguang Xu, S Kevin Zhou, Olivier Pauly, Sasa Grbic, Thomas Mertelmeier, Julia Wicklein, Anna Jerebko, Weidong Cai, and Dorin Comaniciu. 3d anisotropic hybrid network: Transferring convolutional features from 2d images to 3d anisotropic volumes. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 851–858. Springer, 2018.

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

[20] Jingting Ma, Feng Lin, Stefan Wesarg, and Marius Erdt. A novel bayesian model incorporating deep neural network and statistical shape model for pancreas segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 480–487. Springer, 2018.

[21] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 565–571. IEEE, 2016.

[22] Stanislav Nikolov, Sam Blackwell, Ruheena Mendes, Jeffrey De Fauw, Clemens Meyer, Cían Hughes, Harry Askham,

Bernardino Romera-Paredes, Alan Karthikesalingam, Carlton Chu, et al. Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy. *arXiv preprint arXiv:1809.04430*, 2018.

[23] 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.

[24] Arish A Qazi, Vladimir Pekar, John Kim, Jason Xie, Stephen L Breen, and David A Jaffray. Auto-segmentation of normal and target structures in head and neck ct images: A feature-driven model-based approach. *Medical physics*, 38(11):6160–6170, 2011.

[25] Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jonathon Shlens. Stand-alone self-attention in vision models. *arXiv preprint arXiv:1906.05909*, 2019.

[26] Patrik F Raudaschl, Paolo Zaffino, Gregory C Sharp, Maria Francesca Spadea, Antong Chen, Benoit M Dawant, Thomas Albrecht, Tobias Gass, Christoph Langguth, Marcel Lüthi, et al. Evaluation of segmentation methods on head and neck ct: auto-segmentation challenge 2015. *Medical physics*, 44(5):2020–2036, 2017.

[27] Xuhua Ren, Lei Xiang, Dong Nie, Yeqin Shao, Huan Zhang, Dinggang Shen, and Qian Wang. Interleaved 3d-cnn s for joint segmentation of small-volume structures in head and neck ct images. *Medical physics*, 45(5):2063–2075, 2018.

[28] 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.

[29] Holger R Roth, Le Lu, Nathan Lay, Adam P Harrison, Amal Farag, Andrew Sohn, and Ronald M Summers. Spatial aggregation of holistically-nested convolutional neural networks for automated pancreas localization and segmentation. *Medical image analysis*, 45:94–107, 2018.

[30] 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.

[31] Richard Sims, Aurelie Isambert, Vincent Grégoire, François Bidault, Lydia Fresco, John Sage, John Mills, Jean Bourhis, Dimitri Lefkopoulos, Olivier Commowick, et al. A pre-clinical assessment of an atlas-based automatic segmentation tool for the head and neck. *Radiotherapy and Oncology*, 93(3):474–478, 2009.

[32] Jonathan Sykes. Reflections on the current status of commercial automated segmentation systems in clinical practice. *Journal of medical radiation sciences*, 61(3):131–134, 2014.

[33] 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.

[34] Hao Tang, Chupeng Zhang, and Xiaohui Xie. Nodulenet: Decoupled false positive reductionfor pulmonary nodule detection and segmentation. *arXiv preprint arXiv:1907.11320*, 2019.

[35] David N Teguh, Peter C Levendag, Peter WJ Voet, Abrahim Al-Mamgani, Xiao Han, Theresa K Wolf, Lyndon S Hibbard, Peter Nowak, Hafid Akhiat, Maarten LP Dirkx, et al. Clinical validation of atlas-based auto-segmentation of multiple target volumes and normal tissue (swallowing/mastication) structures in the head and neck. *International Journal of Radiation Oncology* Biology* Physics*, 81(4):950–957, 2011.

[36] David Thomson, Chris Boylan, Tom Liptrot, Adam Aitkenhead, Lip Lee, Beng Yap, Andrew Sykes, Carl Rowbottom, and Nicholas Slevin. Evaluation of an automatic segmentation algorithm for definition of head and neck organs at risk. *Radiation Oncology*, 9(1):173, 2014.

[37] Nuo Tong, Shuiping Gou, Shuyuan Yang, Dan Ruan, and Ke Sheng. Fully automatic multi-organ segmentation for head and neck cancer radiotherapy using shape representation model constrained fully convolutional neural networks. *Medical physics*, 45(10):4558–4567, 2018.

[38] Bram Van Ginneken, Cornelia M Schaefer-Prokop, and Mathias Prokop. Computer-aided diagnosis: how to move from the laboratory to the clinic. *Radiology*, 261(3):719–732, 2011.

[39] 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.

[40] René F Verhaart, Valerio Fortunati, Gerda M Verduijn, Aad Lugt, Theo Walsum, Jifke F Veenland, and Margarethus M Paulides. The relevance of mri for patient modeling in head and neck hyperthermia treatment planning: A comparison of ct and ct-mri based tissue segmentation on simulated temperature. *Medical physics*, 41(12), 2014.

[41] Peter WJ Voet, Maarten LP Dirkx, David N Teguh, Mischa S Hoogeman, Peter C Levendag, and Ben JM Heijmen. Does atlas-based autosegmentation of neck levels require subsequent manual contour editing to avoid risk of severe target underdosage? a dosimetric analysis. *Radiotherapy and Oncology*, 98(3):373–377, 2011.

[42] Christian Wachinger, Karl Fritscher, Greg Sharp, and Polina Golland. Contour-driven atlas-based segmentation. *IEEE transactions on medical imaging*, 34(12):2492–2505, 2015.

[43] Gary V Walker, Musaddiq Awan, Randa Tao, Eugene J Koay, Nicholas S Boehling, Jonathan D Grant, Dean F Sittig, Gary Brandon Gunn, Adam S Garden, Jack Phan, et al. Prospective randomized double-blind study of atlas-based organ-at-risk autosegmentation-assisted radiation planning in head and neck cancer. *Radiotherapy and Oncology*, 112(3):321–325, 2014.

[44] Wenzhe Wang, Qingyu Song, Ruiwei Feng, Tingting Chen, Jintai Chen, Danny Z Chen, and Jian Wu. A fully 3d cascaded framework for pancreas segmentation. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 207–211. IEEE, 2020.

[45] Yan Wang, Yuyin Zhou, Wei Shen, Seyoun Park, Elliot K Fishman, and Alan L Yuille. Abdominal multi-organ seg-

mentation with organ-attention networks and statistical fusion. *Medical image analysis*, 55:88–102, 2019.

[46] Zhensong Wang, Lifang Wei, Li Wang, Yaozong Gao, Wufan Chen, and Dinggang Shen. Hierarchical vertex regression-based segmentation of head and neck ct images for radiotherapy planning. *IEEE Transactions on Image Processing*, 27(2):923–937, 2018.

[47] Xingyu Wu, Jayaram K Udupa, Yubing Tong, Dewey Odhner, Gargi V Pednekar, Charles B Simone II, David McLaughlin, Chavanon Apinorasethkul, Ontida Apinorasethkul, John Lukens, et al. Aar-rt–a system for auto-contouring organs at risk on ct images for radiation therapy planning: Principles, design, and large-scale evaluation on head-and-neck and thoracic cancer cases. *Medical image analysis*, 54:45–62, 2019.

[48] Yingda Xia, Lingxi Xie, Fengze Liu, Zhuotun Zhu, Elliot K Fishman, and Alan L Yuille. Bridging the gap between 2d and 3d organ segmentation with volumetric fusion net. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 445–453. Springer, 2018.

[49] Qihang Yu, Lingxi Xie, Yan Wang, Yuyin Zhou, Elliot K Fishman, and Alan L Yuille. Recurrent saliency transformation network: Incorporating multi-stage visual cues for small organ segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8280–8289, 2018.

[50] Tiezhi Zhang, Yuwei Chi, Elisa Meldolesi, and Di Yan. Automatic delineation of on-line head-and-neck computed tomography images: toward on-line adaptive radiotherapy. *International Journal of Radiation Oncology* Biology* Physics*, 68(2):522–530, 2007.

[51] Ningning Zhao, Nuo Tong, Dan Ruan, and Ke Sheng. Fully automated pancreas segmentation with two-stage 3d convolutional neural networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 201–209. Springer, 2019.

[52] Wei Zhao, Jiancheng Yang, Yingli Sun, Cheng Li, Weilan Wu, Liang Jin, Zhiming Yang, Bingbing Ni, Pan Gao, Peijun Wang, et al. 3d deep learning from ct scans predicts tumor invasiveness of subcentimeter pulmonary adenocarcinomas. *Cancer research*, 78(24):6881–6889, 2018.

[53] Xiangrong Zhou, Takaaki Ito, Ryosuke Takayama, Song Wang, Takeshi Hara, and Hiroshi Fujita. Three-dimensional ct image segmentation by combining 2d fully convolutional network with 3d majority voting. In *Deep Learning and Data Labeling for Medical Applications*, pages 111–120. Springer, 2016.

[54] Yuyin Zhou, Lingxi Xie, Wei Shen, Yan Wang, Elliot K Fishman, and Alan L Yuille. A fixed-point model for pancreas segmentation in abdominal ct scans. In *International conference on medical image computing and computer-assisted intervention*, pages 693–701. Springer, 2017.

[55] Wentao Zhu, Yufang Huang, Hui Tang, Zhen Qian, Nan Du, Wei Fan, and Xiaohui Xie. Anatomynet: Deep 3d squeeze-and-excitation u-nets for fast and fully automated whole-volume anatomical segmentation. *arXiv preprint arXiv:1808.05238*, 2018.

[56] Wentao Zhu, Yufang Huang, Liang Zeng, Xuming Chen, Yong Liu, Zhen Qian, Nan Du, Wei Fan, and Xiaohui Xie. Anatomynet: Deep learning for fast and fully automated whole-volume segmentation of head and neck anatomy. *Medical physics*, 46(2):576–589, 2019.

[57] Zhuotun Zhu, Yingda Xia, Wei Shen, Elliot Fishman, and Alan Yuille. A 3d coarse-to-fine framework for volumetric medical image segmentation. In *2018 International Conference on 3D Vision (3DV)*, pages 682–690. IEEE, 2018.